Evaluating the Validity of Score Interpretation in a Libyan EFL Achievement Test: A Construct-Based Analysis

Hamuda, Mustafa

Department of English Language, University of Tripoli, Libya

m.ibrahim@uot.edu.ly

ABSTRACT

This study evaluates the validity of score interpretations of the Libyan preparatory school English final achievement test. Grounded in Messick's (1989) unified validity framework and Bachman and Palmer's (2010) theoretical framework of language knowledge, the study utilizes qualitative content analysis supported by descriptive frequency statistics to evaluate item-level construct representation. The empirical distribution reveals severe construct underrepresentation and construct-irrelevant variance; the examination is strictly restricted to sentence-level grammar recognition (52%), isolated vocabulary tasks (20%), and rote textbook factual recall (28%). Textual, functional, and sociolinguistic dimensions are completely omitted (0%), entirely excluding productive and interactive communicative skills. Consequently, interpreting these test scores as indicators of communicative language ability is empirically unsupported, presenting a significant potential for negative educational washback. The study highlights a critical misalignment between communicative curricular mandates and actual testing practices, calling for immediate reform toward contextualized, performance-based testing.

Keywords: Test Validity, Construct, Communicative Language Ability, Achievement Tests, Language Knowledge

الملخص

تقيم هذه الدراسة مدى صحة تفسيرات درجات اختبار التحصيل النهائي لمادة اللغة الإنجليزية في مرحلة التعليم الأساسي (الشهادة الإعدادية) في ليبيا. واستناداً إلى الإطار الموحد للصلاحية لـ "ميسيك" (1989) والإطار النظري للمعرفة اللغوية لـ "باكمان وبالمير" (2010)، توظف الدراسة منهج التحليل الكيفي للمحتوى مدعوماً بالإحصاء الوصفي للتكرارات لتقييم مدى تمثيل بنية الاختبار على مستوى المفردات والأسئلة. ويكشف التوزيع التجريبي عن قصور حاد في تمثيل البنية اللغوية المستهدفة، إلى جانب وجود تباين لا علاقة له بهذه البنية؛ إذ يقتصر الامتحان بشكل صارم على تمييز القواعد النحوية على مستوى الجملة بنسبة (52٪)، ومهام المفردات المعزولة بنسبة (20٪)، والاسترجاع الآلي القائم على الحفظ للمعلومات الواقعية من الكتاب المدرسي بنسبة (28٪). في المقابل، غابت الأبعاد النصية والوظيفية

والاجتماعية اللغوية تماماً بنسبة (0٪)، مما أدى إلى استبعاد المهارات التواصلية الإنتاجية والتفاعلية بالكامل. وبناءً على ذلك، فإن تفسير درجات هذا الاختبار كمؤشرات على القدرة اللغوية التواصلية يفتقر إلى الدعم التجريبي، مما ينطوي على احتمالية كبيرة لإحداث أثر تراجعي سلبي على العملية التعليمية. وتسلّط الدراسة الضوء على وجود فجوة وعدم اتساق جوهري بين التوجيهات المنهجية القائمة على التواصل والممارسات الاختبارية الفعلية، وتوصي بضرورة إجراء إصلاح فوري نحو تبني اختبارات سياقية قائمة على الأداء.

الكلمات المفتاحية: صلاحية الاختبار، البنية اللغوية ، القدرة اللغوية التواصلية، اختبارات التحصيل الدراسي، المعرفة اللغوية.

Introduction

Achievement tests play a powerful role in shaping educational practices, influencing not only student outcomes but also instructional priorities and classroom practices. In many educational contexts, test scores are often treated as direct indicators of learners' abilities. However, the legitimacy of such interpretations depends on the extent to which they are supported by coherent theoretical frameworks and empirical evidence.

In the field of foreign language education, this issue is critical. Language ability is a complex, multidimensional construct that involves not only knowledge of linguistic forms but also the ability to use language meaningfully in context. Despite this, many large-scale achievement tests continue to rely on discrete-point formats that prioritize practicality and scoring efficiency over construct representation. This creates a tension between communicative curricular goals and traditional testing practices.

Context: Curriculum Reform and Assessment in Libya

The Libyan Ministry of Education, in collaboration with Garnet Education (UK), introduced new English language materials in the late 1990s. Designed for learners aged 11–17, the curriculum allocated approximately four hours of English instruction per week and included student books, workbooks, teacher's guides, and audio materials.

The new materials explicitly claimed adherence to the principles of Communicative Language Teaching. They emphasized varied tasks, pair

and group work, whole-class interaction, and the development of learners' confidence in using English meaningfully (Frino et al., 2008). Teacher's guides articulated lesson objectives and provided procedural guidance intended to support communicative classroom practices.

Despite these curricular intentions, implementation has been uneven. A key factor repeatedly identified in practitioner reports and empirical studies is the misalignment between communicative curricular objectives and testing practices (Masoud, 2017; Orafi & Borg, 2009; Shihiba, 2011). Final tests continue to rely heavily on selected-response formats assessing discrete linguistic knowledge. This misalignment raises fundamental questions about the legitimacy of score interpretations, particularly when examination results are used for promotion and certification decisions.

Research Gap

Although some Libyan researchers have examined Libyan English achievement tests, they have focused on traditional content validity, predictive validity or statistical reliability (Aghanimi et al., 2020; Mahfoud, 2020; Masoud, 2017; Onaiba, 2024). A review of local literature reveals that no empirical studies have systematically applied contemporary, unified validity frameworks to investigate the interpretive legitimacy of Libyan final examination scores. Earlier approaches to validity tended to treat different forms of validity, such as content validity and criterion validity, as separate categories. However, modern validity theory reconceptualises validity as a unified concept centred on construct validity. In this view, different forms of evidence—such as content relevance, internal structure, and relationships with external criteria—are integrated to support the interpretation and use of test scores. This shift allows researchers to evaluate more realistically whether a test actually represents the underlying construct it is intended to measure

Purpose of the study

The purpose of this study is to evaluate the validity of score interpretations of the Libyan preparatory school English achievement test using Messick's (1989) unified validity framework. Specifically, the study examines whether the test provides sufficient evidence to support interpretations of students' communicative language ability.

Research Objectives:

1. Analyse the extent to which the test items represent the theoretical construct of communicative language competence.
2. Examine the cognitive demands and processing levels elicited by the test items relative to expected language use.
3. Evaluate the internal structural alignment of the examination against its intended curricular objectives.

Therefore, the primary objective of this research is to analyse the internal alignment and construct validity of the assessment. To achieve an analytical focus, the scope of this study is document-based, focusing on evaluating the construct representation within the test instrument itself. Because this study does not include empirical, field-based methods—such as interviews, questionnaires, or classroom observations—the findings delineate the test's internal structural properties rather than representing a comprehensive evaluation of test use or practical washback effects.

Research Questions

This study is guided by the following central research question:

-To what extent are the interpretations of scores from the Libyan preparatory school English achievement test supported by evidentiary validity facets within Messick's (1989) unified framework?

To address this primary inquiry, the following specific sub-questions are examined:

a) Content Relevance and Representativeness: To what extent does the test content structurally represent the targeted construct of communicative language competence as prescribed by the national curriculum?

b) Cognitive Aspect: What levels of cognitive processing and task demands are explicitly targeted by the test items relative to the expected domain of language use?
c) Structural Aspect: How is the internal configuration and item distribution of the examination aligned with the structural priorities and weighting of the intended curricular objectives?

Significance of the Study

This study contributes to the field of language testing by applying a contemporary validity framework to a high-stakes national examination in an under-researched context. It moves beyond traditional approaches to validity by focusing on the legitimacy of score interpretations rather than the technical properties of the test alone.

The findings are expected to provide valuable insights for policymakers, curriculum designers, and test developers in Libya, particularly in addressing the misalignment between communicative language teaching objectives and existing assessment practices. More broadly, the study highlights the importance of aligning testing practices with theoretically grounded constructs of language ability in EFL contexts.

LITERATURE REVIEW

Constructs in Language Testing

In language testing, constructs refer to the theoretical attributes or abilities that tests are intended to measure. Constructs such as communicative competence, fluency, and proficiency serve as the conceptual foundations for test design and score interpretation (Newton & Shaw, 2014). Early validity theory conceptualized constructs as stable psychological traits existing independently within the test taker (Cronbach & Meehl, 1955). From this perspective, tests were designed to measure these underlying traits through carefully selected items.

More recent approaches, informed by socio-cognitive and socio-cultural theories, have challenged this. Contemporary scholars argue that language ability is not a fixed internal trait but a context-dependent capacity that emerges through performance in specific communicative situations

(Fulcher, 2025; Weir, 2005). In this view, constructs must be defined in relation to the tasks, contexts, and cognitive processes involved in real-world language use.

Operationalizing a construct requires explicit links between theoretical definitions and observable performance. In communicative language testing (and teaching), this entails designing tasks that require learners to integrate linguistic, pragmatic, and strategic knowledge in order to achieve communicative goals (Bachman & Palmer, 2010; McNamara, 2000).

Evolving Conceptions of Validity

Early validity theory conceptualized a valid test as an instrument designed to measure its targeted underlying psychological construct or behavioural domain. In this view, validity resided in the test itself. It was classified into three types:

1. Content validity: the content of the test should include a representative sample from the domain targeted by the test.
2. Criterion validity: the students' scores on the test would correlate with their scores on another widely accepted measure of the same ability. There are two types:

   a) Concurrent validity: the test student scores would be used to predict some criterion at the time of the doing the test.

   b) Predictive validity: an attempt to make future predictions of the student performance based on the scores on the test.

   3. Construct validity: construct was considered to be 'psychologically real construct' that would have an independent existence in the test taker head. The test scores would measure that construct.

Messick's Unified Validity Framework

Messick (1989) fundamentally redefined the concept of test validity, presenting it as:

> an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and

> appropriateness of inferences and actions based on test scores or other modes of assessment (p.13).

In Messick's (1989) view, validity became a unified concept rather than a set of separate types of validity. It encompasses various sources of evidence, such as content relevance and criterion relationships, which collectively support the interpretation and use of test scores. These aspects are complementary and together provide evidence for a test's construct validity. Accordingly, construct validity emerged as the central concept in contemporary language testing theory. It resides in the interpretation of test scores rather than being an inherent property of the test itself (Weir, 2005). Consequently, construct validity is multifaceted and requires diverse types of evidence to support legitimate inferences about test-takers' abilities based on their scores.

An important advantage of Messick's (1989) framework is that it enables a more comprehensive evaluation of tests. Rather than examining isolated aspects of validity, the unified model considers multiple sources of evidence supporting the interpretation of test scores. This broader perspective allows researchers to determine whether a test adequately represents the abilities it claims to measure and whether the resulting scores can be appropriately used for educational decision-making.

Messick's (1989) framework identifies several sources of validity evidence, including:

1. Content Evidence: does the test content adequately represent the construct domain? This involves analysing how well the test tasks sample the relevant knowledge, skills, and abilities.
2. Substantive Evidence: this refers to the alignment between the cognitive processes elicited by test items and those required in authentic language use. Such evidence may be obtained through methods such as think-aloud protocols, technological tools, or psychometric procedures designed to investigate cognitive processes (Skaggs, 2023).
3. Structural Evidence: this examines the internal structure of the test and whether relationships among test items align with the theoretical construct being measured.

4. External Evidence: this involves analysing the relationship between test scores and external variables. It includes convergent evidence (correlations with other measures of the same construct) and discriminant evidence (lack of correlation with measures of different constructs).
5. Generalisability Evidence: this concerns the consistency of score interpretations across different populations and contexts.
6. Consequential Evidence: this aspect focuses on the social consequences of test use and includes two components:
    a. Value Implications: does the test reflect an appropriate set of values? All tests are inherently value-laden, and therefore the values embedded in a test should be made explicit.
    b. Social Consequences: what are the intended and unintended effects of test use on individuals, groups, and society? This includes both positive and negative effects.

According to Messick (1996), a complete validity argument should consider all these aspects collectively. Bachman (1990) emphasized this point earlier, stating that "none of these by itself is sufficient to demonstrate the validity of a particular interpretation or use of test scores" (p. 237).

Messick (1996) also identified two major threats to construct validity:

1. Construct underrepresentation: this occurs when a test demands too little of the test taker, leading to an oversimplified measurement of the construct. For example, a language test relying only on multiple-choice items may underrepresent communicative ability by neglecting productive language skills.
2. Construct-irrelevant variance: this occurs when a test includes elements unrelated to the ability being measured.

Despite the influence of Messick's (1989) framework, scholars continue to debate whether the social consequences of tests—such as washback—should be considered evidence of validity (Bachman, 1990; Bachman & Palmer, 1996; Cizek, 2020; Fulcher, 2025; Kane, 2006; Messick, 1996). Although tests cannot be separated from their social context, other contextual factors, such as teachers' practices, may sometimes have stronger social consequences than the tests themselves (Hamuda, 2025).

This multi-causal nature of educational outcomes makes it difficult to attribute social consequences solely to tests.

Nevertheless, this debate does not necessarily weaken Messick's (1989) framework. Rather, it reflects an ongoing discussion regarding the scope of validity. Messick's (1989) framework remains influential because it highlights the broader educational implications of testing. Even when scholars disagree about the extent to which consequences should be incorporated into validity evaluation, the framework continues to provide a comprehensive foundation for examining how tests function within educational systems. In the context of the present study, Messick's (1989) framework provides a useful theoretical foundation for examining whether the content of the Libyan preparatory school English test adequately represents the construct of communicative language ability.

Bachman and Palmer's (2010) Language Ability

To operationalize Messick's (1989) broad framework, this study utilizes Bachman and Palmer's (2010) model of communicative language ability. This framework posits that language use is influenced by personal characteristics, topical knowledge, affective schemata, and language ability. Language ability is explicitly divided into language knowledge and strategic competence.

1. Language Knowledge
Language knowledge is the cognitive information in memory that enables users to create and interpret discourse. It is divided into organisational and pragmatic knowledge.
1.1. Organisational Knowledge
This component controls the formal structural system of a language to produce and comprehend grammatically acceptable sentences and cohesive texts. This can be:

1. Grammatical Knowledge: knowledge of vocabulary, morphology, syntax, and phonology required for formal accuracy.
2. Textual Knowledge: it is about rules for structuring spoken and written texts. It encompasses cohesion (reference, substitution,

lexical cohesion, ellipsis, and conjunction) and rhetorical organisation (rhetorical developments like argumentation, comparison-contrast, turn-taking, and topic nomination).

1.2. Pragmatic Knowledge

It governs the relationship between language forms, intended messages, and the sociocultural context to ensure communicative appropriateness. It is divided into:

1. Functional Knowledge: it interprets the relationship between discourse and the user's communicative intentions across four simultaneous functions:

   A) Ideational: expressing experiences by exchanging views and ideas.

   B) Manipulative: affecting the environment via instrumental (requests), regulatory (rules), or interactional (interpersonal relationships) language use.

   C) Heuristic: extending knowledge through learning or problem-solving.

   D) Imaginative: creating imaginary worlds or using language for humour and literature.

2. Sociolinguistic Knowledge: determines the contextually appropriate use of language according to social constraints. It covers genres, dialects/varieties, registers, idiomatic expressions, and cultural references.

2. Strategic Competence

It is reconceptualised as higher-order metacognitive strategies rather than mere compensatory tools, strategic competence manages problem-solving through three executive processes:

1. Goal Setting: identifying language tasks, selecting specific tasks, and committing to their completion.

2. Appraising: evaluating task feasibility and resource demands alongside the user's own topical and language knowledge.

3. Planning: formulating and selecting an optimal communicative response by drawing specific elements from language and topical knowledge.

Developments in Language Testing

Research in language testing has seen a number of different approaches that cover the practice of designing language tests. This research has expanded our understanding of the different factors and processes which influence performance in language tests (Bachman, 2000). The following is a brief discussion of these approaches.

Discrete Point Tests

This stage was informed by the theory of psychometrics and influenced by the school of structural linguistics whereby language ability comprises four language skills (listening, speaking, reading and writing) and knowledge of the grammatical system, vocabulary and pronunciation (Lado, 1961). As a result, language tests focused on testing separately language skills and components. There was a general tendency to decontextualize skills and components being tested. To test this decontextualized knowledge, item formats of multiple choice questions (MCQ) were considered to be the most appropriate technique (McNamara, 2000).

Pragmatic Language Testing

Oller (1979) introduced what he called 'Unitary Competence Hypothesis' in which he claimed that the learner's language proficiency was essentially a single unitary ability rather than separate skills and components, as advocated by Lado (1961). Language testers needed to concentrate not on testing isolated skills and components, but rather on testing the ability of learners to integrate lexical, grammatical, contextual and pragmatic knowledge in the test performance. Examples of these tests included 'cloze tests (gap-filling reading tests)'. Language testers should also be

concerned with the psycholinguistic processing involved in language use which includes:

1) Online processing of language in its real time (naturalistic speaking, listening, etc.)
2) A pragmatic mapping component, whereby formal knowledge of language is a source of expressing and understanding meaning in its context (McNamara, 2000). This paved the way for communicative language testing, which recognizes that language use is contextual, purposeful, and interactive.

Communicative Language Testing

This period is informed by the new theory of language, that of the Theory of Communicative Competence (Hymes, 1972) which influenced language learning and language testing. Hymes (1972) emphasised the importance of understanding the ability of using language in its social context. Thus, there was a shift of emphasis from focusing on the internal psychological processes to focusing on the external social functional aspects of language use (McNamara, 2000). This marked shift in theory has had a significant influence on language testing. According to Rea-Dickins (2000), the change in language testing has generally manifested itself in three areas:

1. Changes in content: the content of tests became broader to involve skills and sub-skills of listening, speaking, reading and writing.

2. Changes in format: a shift from explicitly language focused tasks such as selecting the correct prepositions, to communicatively contextualised tasks which have a specific communicative purpose in mind such as 'listen to announcements and extract specific points of information'.

3. Changes in marking criteria: rather than focusing on criteria associated with the accurate syntax, lexis, etc., the focus is on a new identified set of criteria, with aspects of communicative language use.

Communicative language tests are characterised by two main features. First, performance tests are carried out when the test taker is engaged in the actual extended act of communication, receptively or productively.

Second, language tests focus on social interaction tasks which the language tester will be likely to perform in real life situations. This has led the examiners to use authentic and real world tasks in their tests (McNamara, 2000). Bachman and Palmer (1996) stressed that in communicative language tests there should be a degree of correspondence between language test performance and language use in order for a given test to serve its purpose.

The Impact of Curricular-Assessment Mismatch on EFL Classrooms

Research asserts that tests are never neutral instruments and inherently possess significant washback effects (Stobart, 2003). A critical issue within international English as a Foreign Language (EFL) context is the profound misalignment that often exists between communicative-based curricular reforms and the traditional assessment frameworks used to evaluate them. While ministries of education might mandate updates towards communicative textbooks and teaching goals, the corresponding national assessment systems frequently lag behind, remaining reliant on cost-effective and easily scored discrete-point multiple-choice examinations (Aghanimi et al., 2020; Alhamami, 2021; Cheng, 2005; Wall & Alderson, 1993). This durable challenge has been specifically documented in Libya, where researchers have identified a distinct mismatch between explicitly stated communicative curricular objectives and actual testing practices (Elabbar, 2011; Masoud, 2017; Orafi & Borg, 2009).

This discrepancy shapes classroom dynamics and pedagogy. When faced with such incongruence, teachers tend to ignore essential but non-tested materials, such as specific writing skills, deliberately narrowing the content of their instruction only to those limited areas amenable to the existing test formats (Chalhoub-Deville, 2009; Cheng, 2005; Orafi & Borg, 2009; Wall & Alderson, 1993; Waer, 2017). However, it is to acknowledge that these washback effects are not uniform; they are mediated and influenced by various contextual factors, including rigid institutional constraints and the diverse educational backgrounds of the teachers themselves (Hamuda, 2025).

The impact extends equally to learner behaviour and strategies. Predictably, learners focus their primary efforts on those specific aspects

deemed most likely to appear on examinations. Consequently, they adopt highly instrumental, test-oriented learning strategies that prioritize memorizing frequently tested discrete-point knowledge over developing the broader, more complex communicative competence that the core curriculum theoretically intends to promote. Ultimately, this pattern reflects how restrictive test formats implicitly and effectively redefine the practical construct of what constitutes language ability within that academic context (Alexander, 2010; Bailey, 1996; Rose, 2009).

Methodology

Research Design

This study adopts an evaluative, qualitative research design based on qualitative content analysis (Schreier, 2012), which is supplemented by a limited descriptive quantitative frequency analysis. The primary data source is the official Libyan preparatory school English final achievement test. The test consists of 56 selected-response items, including multiple-choice and true/false questions. To ensure full transparency, a complete copy of the examination is included in Appendix 1. Messick's (1989) unified validity framework provides the analytical foundation, functioning as an evaluative model to assess whether the test items adequately support the interpretation and use of resulting test scores (Skaggs, 2023).

Coding Framework and Construct Representation

To evaluate construct representation, a deductive content analysis matrix was developed using individual test items as the primary units of analysis. The items were classified into five distinct linguistic dimensions adapted from Bachman and Palmer's (2010) taxonomy of language knowledge alongside a category for construct-irrelevant variance:

1. Grammatical Knowledge (Discrete-Point Sentence Level)**:** Items assessing isolated morphosyntactic features, verb tenses, or vocabulary definitions stripped of extended communicative context.
2. Textual Knowledge (Cohesion/Coherence Level): Items requiring the identification of discourse markers, cohesive ties,

or rhetorical organization linking multiple sentences or paragraphs.

3. Pragmatic/Functional Knowledge (Illocutionary Level): Items requiring the interpretation of speech acts, functional meanings, or communicative intent within a given context.
4. Sociolinguistic Knowledge: Items evaluating responsiveness to register shifts, cultural idioms, dialectal variations, or linguistic naturalness.
5. Construct-Irrelevant Memorization: Items requiring the rote recall of isolated factual details from the textbook narrative rather than actual language proficiency.

Coding Reliability and Stability

Given the single-analyst design of this study, a systematic two-phase intra-coder agreement protocol was implemented to mitigate subjective interpretive drift and ensure coding stability. Following the initial coding pass, a two-day temporal interval was observed to minimize recall bias and mitigate memory effects before the researcher conducted a blind re-coding of the entire 56-item instrument.

The intra-coder consensus reached 93%, with 52 out of 56 items matching perfectly across both coding iterations. The remaining four mismatched items exhibited minor classification ambiguities between passive lexical recognition (categorized under Grammatical Knowledge) and rote textbook factual recall (categorized under Construct-Irrelevant Memorization). These minor discrepancies were successfully resolved through iterative reconciliation against the operational matrices derived from Bachman and Palmer's (2010) framework, establishing full interpretive stability.

Findings

This section presents a systematic item-level analysis of the Libyan preparatory school English final examination in light of Messick's (1989) unified validity framework. The analysis draws on concrete examples from the examination to illustrate how the operationalized construct diverges from the communicative competence construct articulated in the national curriculum. The findings and subsequent discussion are organized

according to Messick's six distinct aspects of validity: content, substantive, structural, generalizability, external, and consequential.

Descriptive Frequency Analysis

The item-level analysis utilized a deductive content analysis approach. An a priori codebook was established based on the core components of communicative language competence outlined in Bachman and Palmer's (2010) model of language ability. Each of the 56 test items was treated as a single unit of analysis and deductively assigned to a primary construct category based on the underlying linguistic and cognitive task required to arrive at a correct response.

To capture the test's characteristics accurately, items were also classified by their degree of task contextualization, distinguishing between decontextualized discrete-point items and situational communicative prompts.

Table 1: Distribution of Test Items by Language Knowledge Category (Bachman & Palmer, 2010)

Major Dimension	Language Knowledge Sub-Category	Number of Items	Percentage	Task Contextualise-tion Type
Organizational Knowledge	A. (Grammar recognition)	29	52%	Isolated/ Discrete-point
	B. (Lexis/ Vocabulary)	11	20%	Isolated/ Discrete-point
	C. Textual Knowledge	0	0%	N/A
Pragmatic Knowledge	A. Functional	0	0%	N/A

	Knowledge			
	B. Sociolinguistic Knowledge	0	0%	N/A
Construct-Irrelevant Variance	Factual Content Recall	16	28%	Decontextualized Memorization
Total		56	100%	

The empirical distribution demonstrates that the test instrument is bounded by two categories: discrete-point grammatical/lexical processing (72%) and construct-irrelevant factual memorization (28%). Textual, functional, and sociolinguistic knowledge dimensions are completely unrepresented (0%), indicating complete construct underrepresentation regarding the communicative competencies mandated by the curriculum policy.

Critical Facet Analysis within Messick's Model

Content Aspect

The content aspect of validity concerns the extent to which test tasks represent the targeted construct domain. The analysis demonstrates that the test operationalizes language ability almost exclusively through decontextualized grammar and vocabulary recognition, which fall strictly under 'Organizational (Grammatical) Knowledge'. For example, vocabulary knowledge is tested through isolated word-meaning or synonym/antonym items:

Q13)…….. means funny pictures.
(A) television (B) cartoons (C) fashion (D) arts

Q20) The opposite of ugly.

(A) beautiful (B) boring (C) new (D) old

These items require learners to identify lexical meanings without embedding them in communicative contexts. While such items may provide limited evidence of receptive lexical knowledge, they fail to test learners' Pragmatic Knowledge—specifically the ability to use vocabulary dynamically to achieve communicative purposes, such as describing experiences, expressing opinions, or negotiating meaning. The complete absence of contextualized tasks indicates severe under-sampling of the communicative domain.
Similarly, grammar is tested through sentence-level structural recognition:

Q35) Many people ……. when Tyson arrived at 10 a.m.
(A) waited (B) were waiting (C) is waiting (D) wait

Q44) What were you ……?
(A) does (B) doing (C) do (D) did

These items test learners' recognition of grammatical forms in isolation. They do not require learners to organize extended discourse (Textual Knowledge) or deploy grammatical structures strategically to achieve communicative objectives (Functional Knowledge). As a result, the test content reflects a narrow operationalization of the construct that underrepresents the functional and pragmatic dimensions of language use emphasized in communicative curricula.

Substantive Aspect

The substantive aspect concerns the alignment between the cognitive processes elicited by test tasks and those involved in real-world communicative language use. Authentic communicative language use involves planning utterances, selecting appropriate forms based on socio-pragmatic context, interpreting meaning in extended discourse, and monitoring comprehension. In contrast, the test items predominantly elicit low-level recognition and recall processes.
For example, the following true/false item requires only factual recall or recognition:

Q6) Lina Fakrum was born in Tripoli.
(A) True (B) False

Q9) Hedgehogs can be eaten by foxes.
(A) True (B) False

Such questions do not require any linguistic processing beyond basic comprehension of a simple sentence and knowledge of factual content. The cognitive demand is minimal and unrelated to communicative language competence.

Similarly, items testing factual recall from course content (e.g., identifying facts about famous people or places) conflate language assessment with rote content memorization, thereby introducing construct-irrelevant variance. Learners' success on such items may reflect their memory of textbook facts rather than their English language ability. The absence of tasks requiring extended comprehension, text-level processing, or real-time interaction indicates a substantial mismatch between the cognitive processes elicited by the test and those involved in authentic language use.

Structural Aspect

The structural aspect examines whether the internal organization and scoring of the test align with the theoretical construct. The test employs a uniform dichotomous scoring model in which all items are scored as either correct or incorrect and appear to be equally weighted. This structure fails to capture qualitative differences in performance and provides no opportunity for partial credit or diagnostic feedback.

For example, a learner who demonstrates partial understanding of tense usage or aspect receives the same score (zero) as a learner who demonstrates no understanding at all. This scoring model obscures important distinctions in learners' language development and undermines the interpretability of scores as indicators of communicative competence. Furthermore, because the test relies entirely on a restricted item format with zero task variety, its internal structure fails to reflect the multidimensional, integrated nature of Bachman and Palmer's (2010) communicative construct.

Generalizability Aspect

Generalizability concerns the extent to which inferences drawn from test performance can be extended across tasks, contexts, and domains of language use. Because the test relies exclusively on multiple-choice and true/false formats, its ability to support generalizations about learners' overall communicative language ability is severely constrained. For example, performance on items such as:

Q29) Why not ……?
(A) to go (B) go (C) goes (D) going

This provides evidence of learners' recognition of a fixed grammatical pattern, but offers no basis for generalizing to learners' ability to use suggestions appropriately in spoken interaction or written communication. The absolute absence of listening, speaking, writing, and extended reading tasks further limits the representativeness of the evidence base. Consequently, the scope of inference is restricted to discrete-point knowledge rather than communicative performance across modalities and contexts.

External Aspect

The external aspect of validity concerns the relationship between test scores and other indicators of the same or related constructs. The present study was limited to a document-based analysis and therefore did not include empirical investigation of the relationship between examination scores and external measures of communicative proficiency. Consequently, external validity evidence could not be directly evaluated.

Nevertheless, the narrow construct representation identified in the content and substantive analyses raises theoretical concerns regarding the extent to which the test scores can legitimately be interpreted as indicators of communicative language ability. If scores reflect only isolated grammatical accuracy, they are unlikely to correlate strongly with external measures of functional, real-world communication.

Consequential Aspect

The consequential aspect concerns the social and educational effects of test use. The dominance of discrete-point items and factual recall questions is likely to produce negative washback on teaching and learning. The test format may encourage instructional practices that prioritize grammatical drilling and memorization over communicative language use. Over time, this washback effect undermines the communicative objectives of the curriculum and narrows the enacted curriculum to what is tested.

Furthermore, because the test is used for high-stakes decisions regarding promotion and certification, the consequences of construct underrepresentation are ethically significant. The test may disadvantage learners whose communicative abilities are not adequately captured through selected-response formats. This misalignment raises concerns about fairness and the legitimacy of decisions made on the basis of test scores.

DISCUSSION

The findings indicate that the interpretation of test scores as indicators of communicative language ability is not supported by sufficient validity evidence. In line with Messick's (1989, 1996) view of validity as an integrated evaluative judgment, this discussion synthesizes evidence across content, substantive, structural, generalizability, external, and consequential aspects to assess the legitimacy of score-based inferences. The findings indicate a structurally weak validity argument underlying the test. The most serious threat to validity is pervasive construct underrepresentation, which fundamentally undermines the claim that test scores reflect overall English language achievement.

The analysis of the test reveals significant shortcomings in its construct validity, particularly when juxtaposed against the communicative goals of the national curriculum and contemporary language testing principles. The test comprises fifty-six questions: 12 true/false and 44 multiple-choice items with four options. A critical observation is the pervasive focus on assessing decontextualized vocabulary and grammar (Organizational Knowledge), alongside the rote memorization of course content. The

limited range of question types—exclusively multiple-choice and true/false—constitutes a significant weakness. While these formats offer advantages in terms of rapid and reliable marking, they simultaneously pose challenges, such as increasing the potential for guessing and necessitating stringent security and monitoring procedures.

From a test construction perspective, the inherent difficulty in creating effective distractors for multiple-choice questions (Hughes, 2003) is evident. This challenge likely contributes to the fact that approximately one-fifth of the test questions are simple dichotomous true/false items, thereby increasing the probability of students answering correctly by mere chance. More critically, this restricted array of question types offers minimal, if any, insight into students' true communicative abilities (Alderson et al., 1995).

A paramount finding of this analysis is the total absence of items measuring Textual, Functional, or Sociolinguistic Knowledge, as well as the fundamental language skills of listening, speaking, and writing. These communicative tasks inherently involve both cognitive and social factors, and what Hymes (1972) termed the "knowledge of language as well as the ability to use language in the social context". When these essential skills and knowledge dimensions are not assessed, it becomes impossible to draw valid inferences about critical aspects of communicative competence, such as pragmatic knowledge, sociolinguistic knowledge, functional knowledge, and strategic competence (Bachman & Palmer, 2010). The test does not operationalize these crucial theoretical constructs, which the course materials are explicitly designed to develop.

Instead, the test format overtly reflects the tradition of discrete-point tests which, as Lado (1961) described, prioritize the assessment of isolated variables like vocabulary and grammatical structures. The exclusion of primary communicative skills and the disproportionate focus on decontextualized vocabulary and grammar severely impede the ability to make meaningful inferences about students' communicative abilities based on their test performance. This precisely exemplifies what Messick (1996) termed construct underrepresentation.

The assessment of grammatical knowledge through true/false and multiple-choice formats, covering topics such as present simple tense, past simple, if-conditionals, comparative forms, and prepositions, further highlights this misalignment. While mastering grammatical structures was historically considered central to language ability in earlier approaches like discrete-point testing—where knowing grammar meant mastering grammatical structures in isolation—this view is incompatible with communicative language teaching. In a communicative framework, knowing grammar extends beyond formal accuracy to encompass the effective use of grammar in communication. Bachman and Palmer (2010) clarify that "grammatical knowledge" enables learners to understand and use formally accurate sentences to express communicative meaning. The teaching materials, which aim to develop students' ability to "notice" grammar rules through exposure and practice, and present grammar both inductively and explicitly, directly support these communicative constructs. Therefore, restricting grammar assessment to true/false and decontextualized multiple-choice items fails to capture students' ability to use grammar effectively in communication, thereby undermining the test's construct validity in this critical area.

Similarly, vocabulary assessment, conducted through true/false and multiple-choice formats (e.g., word-opposite, word-meaning, word-definition), also reflects a discrete-point approach to Organizational Knowledge. While the course materials encourage students to infer word patterns, focus on correct spelling, pronunciation, collocations, and word categories to develop linguistic accuracy, and aim to develop fluency through listening, speaking, and reading, the test constructs appear to focus almost exclusively on recognizing decontextualized vocabulary. This directly aligns with the discrete-point tradition that emphasizes passive knowledge of decontextualized vocabulary and pronunciation. Although testing vocabulary recognition skills through MCQs is generally recommended for certain sub-skills (Alderson et al., 1995; Hughes, 2003), this method alone is insufficient to provide robust evidence of students' ability to use vocabulary communicatively.

A particularly concerning aspect of the test is the inclusion of questions that assess knowledge of factual content from the course book. This implies a strong emphasis by test constructors on memorization, as

learners must recall course material to answer these questions. This reliance on memorization appears to be a systemic issue within Libyan education. However, the ability to memorize facts is not explicitly stated or targeted in the objectives of the teaching materials, and it is highly unlikely to contribute to students' development of language use in authentic communication. This further exacerbates the issue of construct underrepresentation and introduces construct-irrelevant variance insofar as student performance could partly reflect memorization abilities rather than communicative language use.

The test purports to be an achievement test, which inherently implies its purpose is to provide evidence of learners' progress and achievement in alignment with course objectives that are rooted in the communicative language teaching (CLT) approach. This necessitates that test providers consider construct definitions that account for the ability to use language in communicative situations. The constructs measured should ideally encompass communicative language skills for reception, production, and interaction in both oral and written modes.

However, the analysis unequivocally demonstrates that the test has been designed primarily on the principle of testing discrete language components such as grammar and vocabulary. This suggests that language test designers in Libya remain significantly influenced by the 1960s tradition of discrete-point tests. The examination of the test's construct validity reveals significant limitations in its design, not only in terms of the relevant constructs being tested but also concerning the legitimacy of inferences made about students' underlying language abilities based on their scores.

The exclusion of fundamental skills such as listening, speaking, and writing fundamentally jeopardizes the construct validity of the test. Even the components that are tested do not accurately reflect the abilities that the course materials aim to develop. For instance, requiring learners to memorize discrete pieces of course content directly conflicts with the course materials' objective of fostering true language use. This comprehensive exclusion of key communicative skills means that valid

inferences about learners' actual language abilities cannot be reliably drawn.

These findings are consistent with international research documenting tensions between communicative curricula and traditional assessment practices in EFL contexts (Alhamami, 2021; Chalhoub-Deville, 2009; Cheng, 2005; Stobart, 2005; Waer, 2017). However, this study extends prior work by explicitly evaluating the validity of score interpretations using Messick's (1989) unified theoretical framework. The study thus reinforces the argument that validity concerns in high-stakes testing contexts must be addressed not only as technical issues but as matters of educational responsibility. Where test scores are used for high-stakes decisions, the lack of a cohesive validity argument raises serious concerns about the legitimacy and fairness of these decisions.

The present study intentionally prioritized theoretical and construct-based analysis of the examination itself rather than stakeholder perceptions or classroom practices. Although field-based methods such as interviews and classroom observations could provide additional insight into washback effects and cognitive processing, document-based validity analysis remains an established approach in language testing research, particularly in initial evaluations of construct representation.

Conclusion

This study provides clear empirical evidence that the score interpretations derived from the national Libyan preparatory school English final achievement examination lack construct validity when evaluated against Messick's (1989) unified framework and Bachman and Palmer's (2010) taxonomy. The examination architecture is characterized by severe construct underrepresentation, excluding all dimensions of textual, functional, and sociolinguistic knowledge, while introducing substantial construct-irrelevant variance through an excessive reliance on factual textbook memorization.

Consequently, the scores generated by this examination cannot legitimately support inferences regarding a student's communicative language competence. Instead, the test functions as a measure of passive

grammar recognition and rote memorization, creating a profound structural mismatch with the national curriculum.

RECOMMENDATIONS

Based on the findings of this study, several recommendations are proposed:

1. Test Reform**:** Future versions of the preparatory English achievement test should incorporate performance-based tasks that assess listening, speaking, reading, and writing in communicative contexts.
2. Professional Development**:** Test developers, inspectors, and curriculum specialists should receive training in contemporary language assessment principles: construct definition and validity theory.
3. Curriculum–Assessment Alignment: Greater coordination is required between curriculum designers and assessment authorities to ensure that testing practices reflect instructional goals.
4. Comprehensive Validity Profiling: Future research should expand upon the current study's evidential and interpretive scope by employing mixed-method approaches involving classroom observation, stakeholder perspectives, and empirical validation procedures.

References

Aghanimi, Y. A., Alwafi, F. M., & Bannur, F. M. (2020). Investigating language tests' content validity in public schools in Tripoli. *Al-Majalla al-Libiyya li-Ulum al-Ta'lim [Libyan Journal of Educational Sciences]*, *1*, 325–381.

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

Alhamami, M. (2021). Teach it, and then set me free: Saudi EFL learners' voice and choice in grammar pedagogies. *Arab World English Journal*, *12*(3), 256–272. https://doi.org/10.24093/awej/vol12no3.18

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. (2000). Modern language testing at the turn of the century. *Language Testing*, *17*(1), 1–42.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assess- ment in practice*. Oxford University Press.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, *13*(3), 257–279.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.

Chalhoub-Deville, M. (2009). The intersection of test impact, validation, and educational reform policy. *Annual Review of Applied Linguistics*, *29*, 118–131.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge University Press.

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.

Cohen, L., Manion, L., & Morrison, K. (2001). *Research methods in education*. Routledge.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Elabbar, F. (2011). Language testing in Libyan schools: A review of assessment practices. *Journal of Language Testing and Assessment*, *14*(2), 45–63.

Frino, L., Mhochain, R. N., O'Neill, H., & McGarry, F. (2008). *English for Libya: Preparatory 3. Teacher's Book*. Garnet Publishing Ltd.

Fulcher, G. (2025). *Practical language testing*. Routledge.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Hamuda, M. (2025). Teachers as key factors in washback: analysing classroom practices in Libyan preparatory schools. *Faculty of Arts Journal*, *1*(41), 1–20.

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

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp.269–293). Penguin.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.,pp.17–64). Praeger.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. McGraw-Hill.

Linn, R. L. (2000). Assessment and accountability. *Educational Researcher*, *29*(2), 4–16.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum* (pp. 83–121). Chicago University Press.

Mahfoud, B. G. (2020). An investigation into the predictive validity of English language assessment at the Technical College of Civil & Meteorology (TCCAM) in Libya. *AL-JAMEAI*, (*32*), 9–27.

Masoud, M. M. (2017). The assessment of English language tests in the General Certificate of Secondary Education in Libya. *Majallat al-Mukhtar lil-Ulum al-Insaniyya [Al-Mukhtar Journal of Humanities]*,*34*(1), 1–19.

McNamara, T. (2000). *Language testing*. Oxford University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256.

Newton, P. A., & Shaw, S. D. (2014). *Validity in educational psychological assessment*. Sage.

Oller, J. (1979). *Language tests at school: A pragmatic approach*. Longman.

Onaiba, A. (2024). Do exam aims and content reflect those of the curriculum? An evaluative study. *Asian Journal of Assessment in Teaching and Learning*, *14*(1), 55–69. https://doi.org/10.37134/ajatel.vol14.1.6.2024

Orafi, S., & Borg, S. (2009). Intentions and realities in implementing communicative curriculum reform. *System*,*37*(2),243–253. https://doi.org/10.1016/j.system.2008.11.004

Schreier, M. (2012). *Qualitative content analysis in practice*. SAGE Publications.

Shihiba, S. (2011). *An investigation of Libyan EFL teachers' conceptions of the communicative learner-centred app-roach in relation to their*

implementation of an English language curriculum innovation in secondary schools [Doctoral dissertation, Durham University]. Durham Research Online

Skaggs, G. (2023). *Test development and validation*. SAGE Publications.

Stobart, G. (2003). The impact of assessment: Intended and unintended consequences. *Assessment in Education: Principles, Policy & Practice*, *10*(2), 139–140.

Waer, H. (2017). Washback from English language tests for primary stage on language teaching and learning: An Egyptian perspective. *Majallat Al-Dirasāt Al-Tarbawiyyah wa Al-Insāniyyah [Journal of educational and human studies]*, *9*(4) ,371–402. https://doi.org/10.12816/0053050

Wall, D., & Alderson, C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing*, *10*(1), 41–69. https://doi.org/10.1177/026553229301000103

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.