Post-graduate Studies and Training

Department of English Applied Linguistics Program

An Investigation of Validity of EFL Writing Assessment to Evaluate EFL Learner's Writing Competence at the Faculty of Education, Sabratha University

A Thesis Submitted in Partial Fulfillment for the Requirements of a Master's Degree in Applied Linguistics

Submitted by:

Asma Mohammed AlAswad

Supervised by:

Dr. Ahmed Alesawe

Academic Year 2025

Abstract

This study investigated the validity of English as a Foreign Language (EFL) writing assessments used to evaluate learner competence at the Faculty of Education-Zulton at Sabratha University. Addressing a notable gap in the Libyan higher education context, the research holds significance by providing empirical evidence on the appropriateness of current assessment practices and offering insights for their improvement. The study aimed to determine if current assessment tools accurately measure learners' writing skills and to investigate the procedures and techniques teachers employ to ensure validity. A mixed-methods sequential design was adopted, utilizing questionnaires administered to a convenience sample of 6 teachers and 30 students, followed by a document analysis of ten previous writing test papers. Data were analyzed using descriptive statistics (SPSS) for the quantitative questionnaire results and a combination of an analytical rubric and thematic analysis for the qualitative data from the writing tests. The findings revealed that while the analytical rubrics used by teachers allowed for a multi-dimensional and reasonably accurate evaluation, significant issues persist. These include a persistent weakness in students' mechanical skills, variability in performance across different tasks, and a critical gap in assessment literacy, where students are largely unaware of the evaluation criteria. Furthermore, a disconnect was observed between frequent in-class assessment and limited out-of-class writing practice, and feedback often resulted in only superficial revisions. The study concludes by recommending the enhancement of assessment validity through increased transparency by sharing rubrics, incorporating more authentic writing tasks, and strengthening the feedback loop to promote deeper engagement and more substantive skill development.

ملخص الدراسة

هدفت هذه الدراسة إلى تقصي صلاحية تقييمات كتابة اللغة الإنجليزية كلغة أجنبية المستخدمة لتقويم كفاءة المتعلمين في كلية التربية زلطن بجامعة صبراتة، تكمن أهمية الدراسة في سدها فجوة بحثية ملحوظة في سياق التعليم العالي الليبي، حيث تقدم أدلة تجريبية حول مدى ملاءمة ممارسات التقييم الحالية ورؤى لتحسينها. سعت الدراسة إلى تحديد ما إذا كانت أدوات التقييم تقيس بدقة مهارات الكتابة لدى المتعلمين، واستكشاف الإجراءات والتقنيات التي يستخدمها المعلمون لضمان هذه الصلاحية، تم اعتماد المنهج البحثي المختلط بتصميم تتابعي، حيث استُخدمت استبيانات وُزعت على عينة قصدية مكونة من ستة معلمين وثلاثون طالبًا، تبعها تحليل وثائقي لعشر أوراق اختبار كتابة سابقة، حُللت البيانات الكمية باستخدام الإحصاء الوصفي (SPSS)، بينما تم تحليل البيانات النوعية باستخدام مقياس تقييم تحليلي (Rubric) وتحليل موضوعي، أظهرت النتائج أنه على الرغم من أن مقاييس التقييم التحليلية المستخدمة توفر تقييمًا متعدد الأبعاد ودقيقًا إلى حد معقول، إلا أنه توجد تحديات كبيرة، أبرزها ضعف مستمر في المهارات الميكانيكية لدى الطلاب، وتباين في الأداء بين المهام المختلفة، ووجود فجوة حرجة في "الوعي التقييمي" لدى الطلاب الذين يجهلون معايير التقييم، كما لوحظ انفصال بين التقييم المتكرر والممارسة الكتابية المحدودة خارج الفصل، وأن التغذية الراجعة غالبًا ما تؤدي إلى مراجعات سطحية، توصي الدراسة بتعزيز صلاحية التقييم عبر زيادة الشفافية بمشاركة مقاييس التقييم مع الطلاب، ودمج مهام كتابية أكثر واقعية، وتقوية حلقة التغذية الراجعة لتعزيز المشاركة الفعالة وتنمية المهارات بشكل أعمق.

Declaration of Originality

This is to certify, that the research paper submitted by me is solely an outcome of my independent and original work. I have duly acknowledged all the sources from which the ideas and extracts have been taken. The project is free from any plagiarism and has not been submitted elsewhere for publication for a degree.

Name of the Researcher: …….

Signature: …….

Dedication

To my parents, my kids and my dear husband who keep supports me .

Acknowledgements

I would like to thank the following people without whom I would not have been able to complete this research, and without whom I would not have made it through my Master Degree!

My supervisor Dr. Ahmed Alesawe for his enduring support and much appreciated advice throughout my thesis. Without his invaluable guidance, this project would not have been possible.

All the teachers and professors who taught and supported me at Zawia University. All the participants who gave me their time to complete the questionnaire and who contributed so thoroughly through their further comments. My gratitude is also expressed to my friends who supported me so positively and always made me feel confident in my abilities.

And my great gratitude to my family, and my husband for all their support, patience and motivation they showed me during my post graduate studies journey.

Table of Contents

List of Tables

List of Abbreviations

Abbreviation	Full Term
APA	American Psychological Association
BERA	British Educational Research Association
CEFR	Common European Framework of Reference for Languages
EFL	English as a Foreign Language
ESL	English as a Second Language
L1	First Language
M	Mean
MMA	Mixed-Methods Approach
MMR	Mixed-Methods Research
N	Number (of participants)
SD	Standard Deviation
SPSS	Statistical Package for the Social Sciences

Chapter One: Introduction

1.0 Introduction

This chapter gives a comprehensive overview of the study that investigates the validity of EFL writing assessments methods that are used to evaluate EFL learners' writing competence at Sabratha University, Faculty of Education-Zulton. It provides the background of the study including the relevance of the topic, the statement of the problem, and the main aims of the research. Besides, the chapter outlines the significant of the study followed by a clear description of the research design and structure.

1.1 Background of the Study

Writing is a vital skill for learning English academically. It is normally considered by some student as the most difficult skill to master. Richard and Renandya (2002) state that the difficulty lies in creating and organizing ideas and translating these ideas into a readable text. According to Ioannou-George and Pavlou (2003:68) "writing is difficult especially in a foreign language". According to Stone, Baugh, & Warfield, (2002), writing is a difficult skill to native speakers as well. Not only it is difficult for student, but for teachers as well. That is, the process of assessing students' writing imposes a great challenge to many teachers. In assessing students, the teacher needs to be aware of what component will be assessed. According to Nunan (2004), assessment is a procedure for gathering students' data. It is a continuous process that involves broader scope. That is, there are plenty of methods to collect information about students' progress and performance.

Benmaar (2016) mention that assessment currently is considered as a great obstacle faced by teachers**.** Black and William (1998a, 1998b) argue that teachers have to guarantee the reliability and validity of their classroom assessment techniques and use these techniques to enhance students' learning.

Thus, the importance of assessment on students' achievements encouraged me to investigate this topic.

1.2 Statement of the Problems

Assessments, as Stiggins and Chappuis (2006) put it, is one of the very crucial topics in school improvement**.** As mentioned by Cooper (2008), there is a broad debate about the assessment methods in higher education. Based on my personal experience during my study years in the secondary school, university as post graduate, the researcher noticed that the assessments methods used in writing were not accurately assessing the ability of students writing competence. In addition, the assessment methods were limited and traditional in that they were centered around the mid and final exams only which lack some other means of assessment such as self and peer assessment. Thus, the researcher tries to shed light on this topic to give an opportunity for teacher to assess writing adequately.

Similar studies in the Libyan context discussed the topic of assessments such as those by El- aswad (2002), Abdul Rahman (2011), Tantani (2012), and Warayet (2013) who discussed assessments on other aspects other than writing. Very similar research conducted by Waragh (2016) who studied the assessment methods used by tutors to test students writing. However, the writing assessments still quite limited (Shihiba ,2011). This means that further studies are still needed to thoroughly investigate the writing assessment in order to pin point the best suitable methods which, in turn, would help in developing the learning of writing skill.

1.3 Aims of the Study

This study the aims to:

1. Explore whether EFL writing assessments at Faculty of Education -Zulton, Sabratha University is valid to test learners' writing.

2. Investigate the procedures and the techniques that are used in order to ensure the validity of the EFL writing assessments.

1.4 Research Questions

According to what have been mentioned earlier, the question that this research is trying to answer are:

Q1. Do EFL learners writing skills accurately assessed using the right EFL writing assessment tools?

Q2. What procedures and techniques do EFL teachers use to ensure the validity of writing assessments?

1.5 Significance of the Study

1.5.1 Theoretical Significance

This research makes substantial theoretical contributions to the field of language assessment and EFL pedagogy. It addresses a notable gap in the scholarly literature concerning assessment validity within the Libyan higher education context, an area that remains considerably underexplored. By investigating the theoretical underpinnings of writing assessment practices, this study extends current understanding of how validity frameworks operate in non-Western educational settings, thereby enriching the global discourse on language assessment.

Moreover, the research responds to scholarly concerns raised by Shihiba (2011) regarding the limited empirical investigation of writing assessment in Libya. It builds upon previous studies in the region while specifically focusing on validity—a fundamental construct in educational measurement. This focus contributes to assessment theory by examining how validity principles are conceptualized and implemented in EFL contexts where English serves as a foreign rather than second language. The findings may offer theoretical insights into the relationship between assessment design, validity evidence, and the accurate measurement of writing competence across diverse educational environments.

1.5.2 Practical Significance

The practical implications of this study are multifaceted and directly relevant to educational stakeholders. For instructors at Sabratha University and similar institutions, the research provides evidence-based insights into current assessment practices, highlighting strengths and identifying areas requiring enhancement. These findings can inform immediate improvements in assessment design and implementation, enabling teachers to select and employ methods that more accurately evaluate student writing

abilities.

Additionally, this investigation offers actionable guidance for curriculum developers and educational administrators seeking to establish quality assurance mechanisms in writing assessment. By documenting the techniques instructors currently use to ensure validity, the study creates a knowledge base that can support professional development initiatives and institutional policy formulation.

From a student-centered perspective, the research has significant implications for learner experiences and outcomes. When assessment methods validly measure writing competence, students benefit from more accurate feedback regarding their strengths and developmental needs. This, in turn, facilitates targeted learning strategies and potentially enhances motivation. Ultimately, improved assessment validity contributes to creating fairer, more transparent evaluation systems that better support students' linguistic growth and academic success in EFL writing.

1.6 Methodology

This study emphasizes the variety of assessment approaches in the field of EFL teaching and learning writing competence. It provides deep insight into the validity and effectiveness of the feedback on the learners‟ achievement regarding the appropriate use of the English language structures in writing which is the concern of many EFL teachers especially at the Faculty of Education at Sabratha University. To be precise, this research sheds light on peer review as one of assessment strategies that is not widely applied at the university context.

Enhancing in-class evaluation by encouraging peer review would result in significant growth in EFL learners' writing correctness, more precisely, on how students use and gain from the writing skills information they have acquired during their university education. It should be emphasized that this technique is supplemental in nature, offering students instantaneous and flexible input rather than a replacement for the teacher's corrective comments. It would be extremely beneficial for EFL practitioners to review this study in order to obtain comprehensive knowledge on creating well-organized peer assessment activities.

This study adopted a mixed-methods approach, combining both quantitative and

Qualitative research methods to provide a comprehensive and nuanced understanding of the research problem. According to Creswell and Clark (2018), mixed-methods research is particularly useful when neither quantitative nor qualitative approaches alone are sufficient to fully understand the complexity of research issue. By integrating both, researchers can benefit from the breadth offered by statistical data and the depth provided by detailed qualitative insights. In the first phase, quantitative data were collected using a close-ended questionnaire administered to both teachers and students. Close-ended questions allow for the collection of standardized responses that can be easily quantified and analyzed statistically **(**Dörnyei, 2007). This method helps to identify general trends, perceptions, and attitudes across a larger sample, making the findings more generalizable. Following the quantitative phase, qualitative data were gathered through a document analysis of previous writing exam papers, specifically from the academic years 2020 to 2022. Document analysis involves systematically reviewing existing documents to extract meaningful information (Bowen, 2009).

It is particularly effective for studying assessment practices over time because it provides real- world evidence of what students were expected to achieve and how they were evaluated. Using this sequential design in which quantitative first then qualitative, allows the study to first map the general patterns through numerical data and then explain or interpret these findings in greater depth through detailed textual analysis. This approach strengthens the validity of the study by offering multiple sources of evidence and provides a more complete understanding of how writing assessments which are perceived and practiced.

1.7 Ethical consideration

This study strictly adhered to ethical research standards to ensure the rights, dignity, and confidentiality of all the participants. Participation in the study was entirely voluntary, and informed consent was obtained from all participants before data collection began. Participants were clearly informed about the purpose of the study, the nature of their involvement, and their right to withdraw at any point without any negative consequences. Anonymity and confidentiality were maintained by coding responses and securely storing all collected data. Furthermore, the data collected were used exclusively for academic purposes and not shared with any third parties. Ethical approval for conducting the research was obtained from the relevant institutional review

board prior to commencing the study, ensuring that the research process complied with international ethical guidelines such as those outlined by the American Psychological Association. (APA, 2017).

1.8 Research Structure

This research is organized into six chapters. Chapter one provides a background of the study outlining the research background, problem statement, aims of the study, and significance of the study. Chapter two presents a review of the relevant literature, offering a theoretical and empirical foundation for the research questions. Chapter three describes the research methodology, detailing the design, instruments, data collection procedures, and methods of analysis. Chapter four focuses on the presentation and analysis of the collected data, using both quantitative and qualitative techniques. Chapter five discusses the findings in relation to the research questions and the existing literature, interpreting the results and highlighting key insights. Finally, Chapter six offers conclusions based on the study's outcomes, along with recommendations for future research and practical implications.

Chapter Two: Literature Review

2.0 Introduction

This chapter presents a comprehensive review of the literature on the assessment of English as a Foreign Language (EFL) writing, focusing on the theoretical foundations, assessment methods, and issues of validity. It begins with an exploration of the theoretical frameworks that have shaped writing instruction and assessment practices over time followed by an overview of various assessment types used in EFL contexts. Key considerations of reliability and validity are discussed in relation to writing assessment methods, emphasizing the importance of ensuring accurate evaluation. The chapter further examines the factors influencing writing performance and reviews empirical studies on the validity of writing assessments, including those conducted within the Libyan educational context. By synthesizing these areas, the chapter provides a critical foundation for understanding the challenges and opportunities in assessing EFL writing competence, setting the stage for the study's investigation into the effectiveness of current practices at Sabratha University.

2.1 Writing Skill

Writing skills are the skills you use to write effectively and succinctly. Writing is the fourth skill in language learning. It involves many manual skills for instance holding a note book properly, handy coordination and copying the letters correctly. Writing is expected to be logical, properly organized speech vanishes but writing can be preserved read and read. In writing appropriate words and context in which they can be used is necessary. Writing is the most powerful medium of human communication that not only involves just a graphical representation of speech but development and presentation of thought into words in a meaningful form and to mentally interact with the message According to Dorothy and Carlos, writing is an important form of communication in day-to-day life, but it is especially important in high school and college.

Writing also one od the most difficult skills to master in both a first language and a second language. Students can find it challenging to find ideas to include their writing, and each culture has its own style for organizing academic writing. Writing relies on vocabulary, grammar, and semantics with the added dependency of a system of sign or

symbols. Good writing skills are needed for all students in order to accomplish their educational and employable requirements. Writing skill which includes all knowledge and abilities to express one idea through the written word.

The act of writing is personal in which writers transfer ideas prompts into topics (Lyons, 1990). In academic disciplines, writing is considered as one of the most important language skills. Most of the academic tasks, mainly graduation projects and exams, are done through writing. Besides, writing helps to clarify students' thinking on a topic, understand what they have read, or remember what they heard in class. For instance, prior knowledge could be activated through free writing before reading a topic; annotating a text while reading enables students to identify important ideas, etc.

Writing is also a complex one in that it integrates many other skills. Thus, the teaching and assessment of this skill needs careful preparation. One reason of this complexity is due to the fact that writing draws on background knowledge and mental processes. To produce a good piece of writing, students have to include several kinds of knowledge such as knowledge of the content in which students conduct a memory search and retrieve prior knowledge and experience, and knowledge of conventions of writing. This inclusion of knowledge leads to some different types of writing. Moreover, besides the linguistic and textual knowledge, metaknowledge about writing such as considering the educational context where students learn to write as well as minimizing the effect of first language (L1) are also considered as other factors that could add to the complexity of writing.

2.1.1 Micro and macro skills of writing

In order for the teachers and the learners alike to understand the complexity of writing mentioned above, Hidri (2020) classified the writing skill into two main categories, micro and macro skills. Micro skill is concerned with the language form and structure. It aims at helping learners with producing correct English graphemes and spelling, using proper word choice and word order, and applying correct grammar rules like tense and agreement. Macro skills, on the other hand, focus on content and meaning through using rhetorical forms and writing conventions. Connecting ideas and events logically with understanding and expressing meanings are on the scope of macro skills of writing. Besides conveying culturally specific information accurately with the use of various

writing strategies (planning, drafting, revising and feedback.

2.1.2 Components of Writing

Effective writing assessment requires a comprehensive understanding of the key components that constitute proficient written communication. Teachers must evaluate multiple dimensions of student writing to provide holistic and meaningful feedback that supports learning and development (Weigle, 2002). According to Hyland (2003), writing is a complex cognitive activity that involves the integration of various linguistic, rhetorical, and mechanical elements. The following components represent the essential areas that teachers should assess when evaluating student writing.

2.1.2.1 Structure

Structure refers to the overall framework and organization of a written text, including how sentences and paragraphs are arranged to convey meaning effectively (Oshima & Hogue, 2006). A well-structured piece of writing demonstrates logical progression of ideas, appropriate use of topic sentences, supporting details, and effective transitions between sections (Langan, 2010). According to Reid (2000), structure encompasses both macro-level organization (the arrangement of major sections) and micro-level organization (sentence-level coherence). Effective structure enables readers to follow the writer's argument or narrative with ease and clarity (Folse, Muchmore-Vokoun, & Solomon, 2010).

2.1.2.2 Vocabulary

Vocabulary assessment involves evaluating the writer's word choice, range of lexical items, and appropriateness of language for the intended audience and purpose (Nation, 2001). Effective vocabulary use demonstrates precision, variety, and sophistication appropriate to the writing task (Read, 2000). According to Schmitt (2000), vocabulary proficiency includes not only breadth (the number of words known) but also depth (the quality of word knowledge, including collocations and register). Teachers assess whether students use subject-specific terminology accurately, avoid repetition through appropriate synonyms, and select words that enhance clarity and impact (Folse, 2004).

2.1.2.3 Content

Content refers to the substance, relevance, and quality of ideas presented in the writing (White & Arndt, 1991). Assessment of content involves evaluating whether the writer has addressed the topic or prompt adequately, developed ideas with sufficient detail and evidence, and demonstrated critical thinking and originality (Ferris & Hedgcock, 2014). According to Grabe and Kaplan (1996), strong content is characterized by depth of analysis, accuracy of information, and the ability to engage the reader through meaningful and well-supported arguments or narratives. Teachers must consider whether the content is appropriate for the intended audience and achieves the communicative purpose of the writing task (Hyland, 2009).

2.1.2.4 Organization

Organization encompasses the logical arrangement and coherence of ideas within and across paragraphs (Connor, 1996). Effective organization includes clear introductions that establish purpose and context, body paragraphs that develop main ideas systematically, and conclusions that synthesize key points (Oshima & Hogue, 2007). According to Johns (1997), good organization is marked by unity (all sentences relate to the main idea), coherence (ideas flow logically through appropriate transitions and connections), and emphasis (important ideas are strategically positioned for maximum impact). Teachers assess whether the organizational pattern chosen—such as chronological, compare-contrast, cause-effect, or problem-solution—is appropriate for the writing purpose and executed effectively (Smalley, Ruetten, & Kozyrev, 2012).

2.1.2.5 Mechanics

Mechanics refers to the technical aspects of writing, including grammar, punctuation, spelling, capitalization, and formatting conventions (Truss, 2003). Proficiency in mechanics ensures that writing is clear, professional, and free from distracting errors that impede comprehension (Lunsford & Lunsford, 2008). According to Ferris (2011), while mechanical accuracy alone does not guarantee effective writing, consistent errors in mechanics can significantly undermine a writer's credibility and the reader's ability to focus on content. Teachers assess mechanics to determine whether students demonstrate control over sentence structure, appropriate use of punctuation marks, correct spelling (particularly of academic or technical vocabulary), and adherence to

formatting requirements such as margins, spacing, and citation styles (Hacker & Sommers, 2011).

Understanding these five components enables teachers to conduct comprehensive writing assessments that provide students with targeted feedback across multiple dimensions of their writing performance. As Weigle (2002) emphasizes, effective writing assessment considers all these elements in relation to one another, recognizing that proficient writing requires the successful integration of structure, vocabulary, content, organization, and mechanics.

2.2 Assessment

Assessment is a comprehensive term that encompasses various methods and tools used to evaluate, measure, and document students' academic skills, learning progress, and educational needs (Brown, 2004; Harlen, 2007). In the realm of education, assessment serves multiple purposes: it provides evidence of student learning, informs instructional decisions, and supports the improvement of educational programs (Black & Wiliam, 1998; Stiggins, 2005). According to Brown (2004), assessment is fundamentally intertwined with instruction, serving as an ongoing process that reveals students' capacity to perform learning tasks and achieve educational objectives.

Assessment can take many forms and serves different functions within the educational context. It may be formal or informal, formative or summative, and can include diverse methods such as observations, discussions, projects, portfolios, presentations, and tests (Popham, 2008; McMillan, 2014). The primary goal of assessment is to provide meaningful feedback to both learners and educators to enhance the learning process and improve educational outcomes (Nicol & Macfarlane-Dick, 2006). Effective assessment practices are characterized by their validity, reliability, and ability to authentically measure what students know and can do (Messick, 1989; Wiggins, 1998).

Contemporary perspectives on assessment emphasize its role not merely as a measurement tool but as an integral component of the teaching and learning cycle (Wiliam, 2011). As Shepard (2000) argues, assessment should be viewed as a learning activity that helps students develop metacognitive awareness and self-regulatory skills. This shift from assessment of learning to assessment for learning represents a fundamental reconceptualization of how evaluation functions within educational

settings (Earl, 2003). The relationship between assessment and instruction is reciprocal: assessment informs teaching practices, while effective instruction shapes what and how students learn, which in turn is reflected in assessment outcomes (Heritage, 2010).

Understanding the various types of assessment and their appropriate applications is essential for educators seeking to optimize student learning. The following section provides a detailed examination of the primary types of assessment used in educational contexts.

2.2.1 Types of Assessment

With regard to the categories of evaluation, Brown (2004) claims that there are four main kinds of assessments that are applied in the classroom with varying goals in mind: formative assessment, summative assessment, formal assessment, and informal assessment.

2.2.1.1 Formative Assessment

Is often understood to be a continuous process that occurs within the teaching and learning environment. The main goal of formative assessment is to provide students with quick feedback so they may improve their learning (Black & Wiliam, 1998). According to Andrade and Cizek (2010), it encompasses all of the actions that educators and/or students engage in that give information that may be utilized as feedback to change the lessons they are teaching and learning. Teachers do not assign a final mark to students' work in this kind of evaluation. Instead, they send them feedback about their learning progress (Irons, 2008). Harlen and James (1997) emphasize that formative assessment is fundamentally concerned with supporting learning in progress rather than measuring achievement at a fixed point in time.

Summative Assessment is the second type. According to Brown (2003), this type of assessment aims to assess or enumerate the knowledge that a student has acquired; this usually happens at the conclusion of a course or a unit of teaching. This kind is usually created ahead of time to allow students enough time to be ready for the test. Scriven (1967) originally distinguished summative assessment as evaluation designed to determine the overall effectiveness of a program or course. Taras (2005) notes that summative assessment serves an accountability function, providing evidence of student

achievement that can be reported to stakeholders.

2.2.1.2 Summative Assessment

Is the second type. According to Brown (2003), this type of assessment aims to assess or enumerate the knowledge that a student has acquired; this usually happens at the conclusion of a course or a unit of teaching. This kind is usually created ahead of time to allow students enough time to be ready for the test. Scriven (1967) originally distinguished summative assessment as evaluation designed to determine the overall effectiveness of a program or course. Taras (2005) notes that summative assessment serves an accountability function, providing evidence of student achievement that can be reported to stakeholders.

2.2.1.3 Formal Assessment

Is the third type of assessment by which organized, methodical approaches to evaluation are used to gauge pupils' language proficiency (Brown, 2004). Students understand that their work will be evaluated when they participate in formal assessments. Examinations and diagnostic procedures are illustrations of formal evaluation used in the classroom to gauge how much progress the pupils have made. These forms feature structured grading procedures and are administered in a standardized manner (Brown, 2003).

2.2.1.4 Informal Assessment

Is the final type, which refers to any type of spontaneous feedback or comment that the teacher provides on the student's work (Brown, 2004). It may contain phrases like 'good work,' 'continue,' and so forth. More significantly, the purpose of the teacher's informal evaluation, which takes place during teaching, is not to determine the outcome or grade of the students' work. For instance, it could be the teachers' brief remarks on the students' papers and their suggestions for improvement, such as how to write an argumentative essay more effectively (Irons, 2008).

2.2.2 Principles of Assessment

A number of factors are taken into account in order to satisfy the efficacy of the assessment activity: authenticity, wash back, validity, reliability, and practicality (Brown, 2004).

1. **Reliability:** when the outcomes of the evaluation tools remain consistent across several scenarios, they are considered dependable. Thus, if the same test is given to the same students or matched students on two different occasions, the test should yield similar results, (Brown, 2004). For instance, the outcomes will be the same if the teacher assigns various tasks to his pupils to complete and then delivers the same assignments to the same students ten days later.

2. **Validity:** a valid assessment is one which measures that it is supposed to measure. As stated, by McAlpine (2002), this approach id deemed acceptable when the kind of evaluation utilized in the classroom evaluates the appropriate skill intended to be tested. To arrange for an evaluation to be considered valid, it must center around the lesson's objectives. For instance, oral ability alone should be required for an oral production test.

3. **Practicality:** this idea refers to determining the practicality of a test. Initially, a test is considered practical if it does not take a lot of time to complete. In other words, it should not be either too long or too short. It should not be very exorbitant or pricey. The ease of scoring is another characteristic of practicality. Put another way, the scorer needs to decide which scoring method is best for the particular test. In addition, appropriate and beneficial settings for test administration are necessary for the exam to be practical, Brown. (2004). In order to assist teachers in selecting the most appropriate teaching and assessment methods for each student's level, test results must provide precise descriptions of the students' proficiency and level.

4. **Authenticity:** this implies that assessments ought to reflect actual situations. In other words, all kinds of evaluation instruments ought to get the student ready to perform appropriately in the intended culture. Additionally, the pieces should be contextualized, and the themes covered should be engaging.

5. **Washback**: The concept of washback, as articulated by Brown and Abeywickrama (2010), refers to the influence that assessment practices exert on both pedagogical approaches and learning processes, encompassing both intentional and unintentional curricular modifications. Effective washback manifests through several key characteristics: it shapes instructional

methodologies and content delivery in constructive ways, guides learners toward meaningful engagement with material, provides adequate preparation opportunities for students, delivers feedback mechanisms that facilitate language acquisition, emphasizes formative rather than summative evaluation principles, and establishes optimal conditions for learners to demonstrate their capabilities. The reciprocal nature of positive washback creates a beneficial cycle wherein enhanced student motivation subsequently influences teacher effectiveness, thereby cultivating an improved classroom dynamic (Linville, 2011, p.15). Consequently, educators must critically evaluate their assessment designs through a comprehensive framework that considers practicality, reliability, validity, authenticity, and the potential for generating constructive washback effects. This reflective practice ensures that assessment tools serve not merely as measurement instruments but as catalysts for meaningful educational advancement.

2.3 Writing Assessment

Writing assessment represents a specialized domain within the broader field of language evaluation, characterized by its complexity and multifaceted nature. Unlike other language skills that may be assessed more objectively, writing evaluation requires careful consideration of multiple linguistic, cognitive, and communicative dimensions. Hughes (2003) emphasizes that assessing writing competence involves examining not merely the final product but also understanding the processes writers employ to construct meaning and communicate effectively in written form.

The assessment of writing in EFL contexts presents unique challenges that distinguish it from evaluating other language skills. Weigle (2002) notes that writing assessment must account for various factors including linguistic accuracy, organizational coherence, rhetorical effectiveness, and the writer's ability to adapt their discourse to different audiences and purposes. Furthermore, Hyland (2003) argues that effective writing assessment should recognize writing as a social practice embedded within specific contexts, rather than viewing it as an isolated demonstration of grammatical knowledge.

Contemporary approaches to writing assessment have evolved considerably from

traditional methods that focused predominantly on error identification and grammatical correctness. Current assessment frameworks acknowledge the recursive nature of writing and the importance of evaluating both process and product dimensions. As Cumming (2001) observes, meaningful writing assessment should provide learners with insights into their developmental trajectory while simultaneously offering instructors diagnostic information to inform pedagogical interventions.

2.3.1 Types of Writing Assessment

Writing assessment encompasses diverse approaches, each serving distinct pedagogical purposes and offering unique insights into learner competence. Understanding these various types enables educators to select assessment methods that align with their instructional objectives and student needs.

2.3.1.1 Direct Writing Assessment

Direct writing assessment involves evaluating actual writing samples produced by learners in response to specific prompts or tasks. This approach, as Weigle (2002) explains, provides authentic evidence of a writer's ability to generate, organize, and express ideas in written form. Direct assessment typically occurs under controlled conditions where students compose texts within specified time constraints and according to predetermined guidelines. The advantage of this method lies in its authenticity—it requires students to demonstrate the very skill being assessed rather than relying on proxy measures. However, Hamp-Lyons (1991) cautions that direct assessment can be resource-intensive, requiring substantial time for administration and evaluation, and may be influenced by factors such as prompt selection and rating consistency.

2.3.1.2 Indirect Writing Assessment

Indirect assessment methods evaluate writing ability through multiple-choice tests, error identification exercises, or sentence completion tasks that measure knowledge about writing rather than writing performance itself. Proponents of indirect assessment, such as those discussed in Bachman & Palmer (1996), argue that these methods offer efficiency and objectivity, allowing large-scale evaluation with

standardized scoring procedures. Nevertheless, critics contend that indirect measures fail to capture the complexity of actual writing performance. Huot (2002) argues that such assessments may test metalinguistic knowledge without necessarily reflecting a learner's capacity to produce coherent, purposeful written texts in authentic communicative contexts.

2.3.1.3 Holistic Assessment

Holistic assessment approaches evaluate writing samples as unified wholes, assigning a single overall score that represents the assessor's general impression of quality. White (1985) describes holistic scoring as particularly useful when evaluators need to process large volumes of writing efficiently while maintaining reasonable reliability. Raters using this method consider multiple dimensions simultaneously—content, organization, language use, and mechanics—without assigning separate scores to individual features. The primary strength of holistic assessment lies in its recognition that effective writing represents more than the sum of its parts; it acknowledges the integrated nature of writing competence. However, Weigle (2002) notes that holistic scores provide limited diagnostic information, making it difficult for learners to identify specific areas requiring improvement or for instructors to target particular weaknesses in their teaching.

2.3.1.4 Analytic Assessment

In contrast to holistic methods, analytic assessment employs scoring rubrics that evaluate writing across multiple distinct criteria or dimensions. Hyland (2003) explains that analytic scoring typically addresses separate aspects such as content development, organization and coherence, vocabulary range and accuracy, grammatical control, and mechanical correctness. Each dimension receives an independent score, which may be weighted differently according to the assessment's priorities. This approach offers substantial advantages for instructional purposes, as it generates detailed diagnostic profiles that illuminate specific strengths and weaknesses in student writing. Jacobs et al. (1981) developed one of the most widely adopted analytic scales for ESL composition, demonstrating how systematic attention to distinct writing features can enhance both reliability and instructional utility. The trade-off, as Hamp-Lyons (2003) observes, involves

increased time investment and the potential for raters to lose sight of the text's overall effectiveness when focusing intently on individual components.

2.3.1.5 Portfolio Assessment

Portfolio assessment represents a comprehensive approach that evaluates collections of student writing assembled over extended periods. Hamp-Lyons and Condon (2000) characterize portfolios as purposeful compilations that may include multiple drafts, reflective commentary, and diverse text types, thereby providing a longitudinal perspective on writing development. This method aligns with process-oriented pedagogies by documenting growth over time and acknowledging the recursive nature of writing development. Portfolios enable assessment of a broader range of writing abilities than single-sitting examinations can capture, and they encourage student agency through selection and reflection processes. Furthermore, Elbow & Belanoff (1997) argue that portfolio assessment can reduce test anxiety by distributing evaluation across multiple occasions and allowing revision. Challenges associated with portfolios include standardization difficulties, substantial evaluation time requirements, and questions regarding the authenticity of work completed outside supervised conditions.

2.3.1.6 Performance-Based Assessment

Performance-based writing assessment emphasizes authentic tasks that simulate real-world writing situations encountered beyond the classroom. Wiggins (1998) advocates for assessments that require students to complete purposeful writing activities resembling those performed by competent writers in academic, professional, or civic contexts. Such tasks might include composing research reports, crafting persuasive arguments for authentic audiences, or creating genre-specific texts that serve genuine communicative functions. This approach enhances validity by ensuring that assessment tasks mirror the ultimate learning objectives—preparing students to write effectively in contexts that matter to them. McNamara (1996) suggests that performance-based assessment can increase student motivation by demonstrating the practical relevance of writing skills. Implementation challenges include designing appropriately authentic tasks that remain feasible within educational constraints and establishing reliable scoring

procedures for complex, open-ended performances.

2.3.2 The Role of Rubrics in Writing Assessment

Rubrics have emerged as essential tools in contemporary writing assessment, serving as structured frameworks that articulate performance expectations and guide evaluative judgments. Fundamentally, a rubric constitutes a coherent set of criteria used to assess student work, specifying both the dimensions to be evaluated and the standards of performance along a quality continuum (Brookhart, 2013). In writing assessment contexts, rubrics function as bridges connecting instructional objectives, learning activities, and evaluative practices, thereby promoting alignment and transparency throughout the educational process.

2.3.2.1 Defining Characteristics and Components

Effective writing rubrics typically incorporate several key elements that work in concert to support reliable and meaningful assessment. According to Andrade (2000), well-constructed rubrics include clearly defined criteria that identify the specific aspects of writing to be evaluated, such as thesis development, evidence quality, organizational structure, stylistic appropriateness, and linguistic accuracy. Each criterion is accompanied by performance level descriptors that characterize what achievement looks like across a range from exemplary to unsatisfactory. Moskal (2000) emphasizes that these descriptors should employ concrete, observable language rather than vague qualitative terms, enabling both assessors and students to recognize the distinguishing features of different performance levels.

The structure of rubrics may vary considerably depending on their intended purpose and the assessment context. Analytic rubrics provide separate scores for each criterion, yielding detailed diagnostic profiles that illuminate specific competencies and deficiencies (Nitko & Brookhart, 2007). Conversely, holistic rubrics generate single overall scores based on general impressions of quality, trading specificity for efficiency (Spandel, 2012). Popham (2008) notes that the choice between these formats should be guided by whether the primary goal is summative evaluation,

formative feedback, or some combination thereof.

2.3.2.2 Functions in Promoting Assessment Validity and Reliability

Rubrics play a crucial role in enhancing both the validity and reliability of writing assessment—two fundamental qualities that determine whether evaluations are meaningful and consistent. Regarding validity, rubrics help ensure that assessments measure what they purport to measure by making evaluation criteria explicit and directly tied to learning objectives. Moskal and Leydens (2000) argue that when rubric criteria genuinely reflect the valued dimensions of writing competence, assessment results provide authentic evidence of student achievement rather than reflecting tangential or construct-irrelevant factors. This content validity is further strengthened when rubrics are developed through careful consideration of disciplinary expectations and expert judgment regarding what constitutes proficient writing in particular contexts.

With respect to reliability, rubrics reduce subjectivity and enhance consistency across raters, occasions, and writing samples. Jonsson and Svingby (2007), in their systematic review of rubric research, found substantial evidence that well-designed rubrics improve inter-rater reliability by providing a common reference framework that guides evaluative judgments. When multiple assessors use the same rubric with clear descriptors, their scores tend to converge, indicating that they are applying consistent standards. Similarly, rubrics can enhance intra-rater reliability—an individual assessor's consistency over time—by serving as stable reference points that counteract the influence of fatigue, mood, or other extraneous variables that might otherwise affect scoring (Arter & McTighe, 2001).

2.3.2.3 Supporting Formative Learning and Student Development

Beyond their evaluative functions, rubrics serve powerful pedagogical purposes that directly support student learning and writing development. When shared with students before they begin writing tasks, rubrics function as instructional tools that clarify expectations and guide effort allocation. Andrade (2001) found that students who received rubrics in advance demonstrated improved performance, suggesting that explicit criteria help learners understand quality standards and make strategic decisions about where to focus their attention during composing processes.

Rubrics also facilitate meaningful feedback that promotes growth. Rather than receiving vague comments like "needs improvement," students can consult rubric descriptors to understand precisely which aspects of their writing require attention and what improved performance would entail. Panadero and Jonsson (2013) emphasize that rubrics transform feedback from subjective judgment into actionable guidance, empowering students to engage in targeted revision and self-regulation. This transparency is particularly valuable in EFL contexts where linguistic and cultural differences may create uncertainty about academic writing conventions.

Furthermore, rubrics support the development of self-assessment and peer-assessment capabilities—metacognitive skills essential for autonomous learning. When students internalize rubric criteria through repeated exposure and application, they become better equipped to evaluate their own writing critically and revise strategically (Andrade & Du, 2005). Ross (2006) documents that training students to use rubrics for self-assessment enhances both the accuracy of their self-evaluations and the quality of their written products. Similarly, structured peer review activities guided by rubrics enable students to learn from examining classmates' work while developing their own critical reading and evaluative capacities (Topping, 2009).

2.3.2.4 Challenges and Considerations in Rubric Implementation

Despite their considerable benefits, rubrics are not without limitations and implementation challenges that educators must navigate thoughtfully. One significant concern involves the potential for rubrics to constrain creativity and encourage formulaic writing. Kohn (2006) cautions that overly prescriptive rubrics may lead students to focus on satisfying checklist requirements rather than engaging authentically with ideas or taking stylistic risks. This concern is particularly relevant when rubrics emphasize superficial features at the expense of deeper qualities like originality, voice, or rhetorical effectiveness.

Additionally, developing high-quality rubrics requires substantial expertise, time, and iterative refinement. Wilson (2006) notes that poorly constructed rubrics with vague descriptors or misaligned criteria can actually undermine assessment quality

rather than enhancing it. Effective rubrics must strike delicate balances—being specific enough to guide judgment without becoming reductive, comprehensive enough to capture important dimensions without becoming unwieldy, and stable enough to ensure consistency without ignoring contextual variations in writing tasks.

Cultural and linguistic considerations present additional complexities in EFL assessment contexts. Rubric criteria developed in native English-speaking contexts may not fully account for the particular challenges faced by language learners or may privilege certain rhetorical traditions over others. Crusan (2010) emphasizes the importance of adapting rubrics to reflect the specific competencies and developmental trajectories relevant to EFL learners, ensuring that assessment standards are appropriately challenging yet attainable.

2.3.2.5 Best Practices for Effective Rubric Use

Research and practitioner wisdom suggest several principles for maximizing rubrics' contribution to effective writing assessment. First, rubrics should be developed collaboratively, involving both instructors and students when feasible, to ensure that criteria reflect shared understanding of quality and are perceived as legitimate (Popham, 2008). Second, rubrics require ongoing validation through trial applications, revision based on user feedback, and periodic review to ensure continued alignment with evolving learning objectives (Tierney & Simon, 2004). Third, educators should provide explicit instruction in rubric interpretation and use, not assuming that criteria are self-explanatory. Panadero and Jonsson (2013) demonstrate that rubrics' effectiveness increases significantly when accompanied by modeling, examples of work at different levels, and opportunities for practice with feedback. Fourth, while rubrics provide valuable structure, they should not entirely replace qualitative feedback and professional judgment. Sadler (2009) argues that assessment represents a skilled practice that cannot be fully automated or reduced to algorithms; rubrics should support rather than substitute for expert evaluation.

Finally, rubrics themselves should be treated as pedagogical artifacts subject to critical examination and improvement. Engaging students in analyzing and even

co-constructing rubrics can deepen their understanding of writing quality while simultaneously enhancing the validity and educational value of assessment instruments (Stiggins, 2001). Through such reflective practices, rubrics can evolve from static checklists into dynamic tools that foster assessment literacy and support continuous improvement in both teaching and learning.

Table (1): Analytic Rubric for Assessing EFL Writing Competence

Criteria	Excellent (5)	Very Good (4)	Good (3)	Fair (2-3)	Poor (2)
Structure					
Vocabulary					
Organization					
Grammar					
Content					

2.4 Validity

The general concept of validity is defined as "the degree to which a test measures what it claims, or purports, to be measuring" (Brown, 1994, p.231). In terms of assessment validity could be defined as "the results obtained from the given measurement procedure objectively reflect the phenomenon the said procedure is intended to measure."(Dobric,2018 p.56). The concept of validity in language testing has evolved over the decades, beginning with early theorists like Lado (1961) and Davies (1968), who focused on aspects such as face validity, content, control of extraneous factors, and empirical insights. Campbell and Fiske (1959) introduced the idea of convergent and discriminate validity, emphasizing that related measures should not. Campbell and Stanely (1966) further differentiated between internal and external validity. A significant advancement came with Bachman (1990), who built on Messick's unified theory of validation from 1989. Bachman identified three key factors that support validity: content, context and criterion validity.

2.4.1 Types of Validity

Validity is considered as an important element in almost all types of assessments.

This means that it has many types as explained below.

2.4.1.1 Content Validity

The degree to which a test or assessment accurately gauges the intended subject matter or skill set is known as content validity. Content validity in writing tests guarantees that the tasks cover all aspects of the writing abilities being assessed, including organization, coherence, argument development, and grammar (Weigle, 2002). A test with high content validity encompasses all essential facets of the ability being measured and is in line with the learning objectives (Brown, 2004). An assessment that is intended to gauge academic writing proficiency but simply consists of multiple-choice questions centered on grammar, for instance, would not be considered content valid since it would not gauge critical abilities like organization and idea development. Expert judgment is required to ensure content validity, whereby linguists or educators examine the test items to ensure that they accurately reflect the writing skills do main (Bachman & Palmer,1996). This procedure aids in avoiding construct irrelevance, which involves evaluating unrelated skills, and construct underrepresentation, which involves leaving out important components of writing proficiency (Messick, 1989).

2.4.1.2 Construct Validity

Construct validity represents the extent to which an assessment instrument accurately measures the theoretical concept it purports to evaluate (Messick, 1989). In the context of writing assessment, this principle ensures that tests capture the essential dimensions of writing ability—including organizational coherence, logical reasoning, and linguistic competence—rather than peripheral factors such as penmanship or test-wiseness strategies (Weigle, 2002). An assessment demonstrating strong construct validity evaluates the comprehensive range of competencies that constitute writing proficiency, moving beyond isolated features like grammatical accuracy or lexical choice to encompass the integrated skills writers employ in authentic contexts (Bachman & Palmer, 1996).

Consider, for example, a writing examination that claims to measure academic composition skills but evaluates only sentence-level grammatical correctness. Such an instrument would exhibit poor construct validity because it fails to assess higher-

order competencies essential to academic writing, particularly the ability to structure arguments coherently and develop ideas systematically. Brown (2004) emphasizes that establishing construct validity requires a multifaceted approach involving empirical investigation, statistical procedures such as factor analysis, and consultation with subject matter experts. This rigorous validation process confirms that assessment outcomes genuinely reflect the target writing construct rather than irrelevant variables or measurement artifacts.

When construct validity is compromised, test results become problematic interpretations of student performance. Scores may misrepresent learners' actual writing capabilities, leading educators to draw inaccurate conclusions about students' strengths and instructional needs. Therefore, maintaining construct validity remains fundamental to developing meaningful assessments that provide authentic evidence of writing proficiency and inform effective pedagogical decisions.

2.4.1.3 Criterion Validity

The degree to which test results align with an external criterion that is recognized to assess. The same skill is known as criterion validity (Brown ,2004). It establishes if a test accurately predicts or corresponds with performance in the real world or with another recognized evaluation. Two categories of criterion validity are frequently distinguished: The degree to which a test corresponds with an evaluation that is already in place and taken concurrent validity (Weigle,2002). For instance, a new writing test has high concurrent validity if it yields results that are comparable to those of a reputable standardized writing exam. The ability of a test to forecast future performance in a related skill is known as predictive validity (Bachman & Palmar,1996). For instance, a student's performance on academic writing assignments in college should be predicted by their performance on a university entrance exam. Researchers compare test results with external metrics, including grades, professional evaluations, or standardized test results, in order to prove criterion validity (Messick,1989). A writing assessment's criterion validity may be compromised if it is intended to gauge student's academic writing proficiency but has little bearing on how well they perform on real assignments.

2.4.1.4 Face Validity

Face validity refers to the extent to which an assessment appears, on its surface, to measure what it claims to assess from the perspective of test-takers, administrators, and other stakeholders who lack specialized psychometric expertise (Mosier, 1947). Unlike other forms of validity that require empirical evidence or statistical analysis, face validity represents a subjective judgment about whether a test seems appropriate, relevant, and credible to those who encounter it (Holden, 2010). Nevo (1985) distinguishes face validity from content validity by emphasizing that while content validity demands systematic expert evaluation of whether test items adequately sample the domain of interest, face validity concerns itself primarily with superficial appearances and user perceptions.

In the context of writing assessment, face validity assumes particular importance because stakeholder acceptance significantly influences test utility and educational impact. When students perceive a writing task as authentic and meaningful—for instance, composing an argumentative essay on a relevant topic rather than completing decontextualized grammar exercises—they are more likely to engage seriously with the assessment and accept results as legitimate indicators of their abilities (Huang, Shear, & Stevenson, 2016). Similarly, instructors and administrators who view an assessment as face valid are more inclined to implement it conscientiously and utilize results for instructional decision-making (Kane, 2013).

Nevo (1985) argues that face validity, though often dismissed as merely cosmetic, serves several pragmatic functions within educational contexts. High face validity can enhance test-taker motivation, reduce anxiety, and increase cooperation during assessment administration. Conversely, assessments lacking face validity may encounter resistance from examinees who question the relevance or fairness of tasks, potentially undermining performance and generating scores that fail to reflect true competence (Anastasi & Urbina, 1997). For example, if an EFL writing assessment requires learners to compose business correspondence despite their academic orientation, students may perceive the task as inappropriate, regardless of its technical psychometric qualities.

However, researchers consistently caution against conflating face validity with more

substantive forms of validity evidence. Messick (1989) emphasizes that an assessment can appear valid without actually measuring the intended construct, while conversely, a test with strong empirical validity might initially seem irrelevant to naive observers. Brown (2004) illustrates this distinction by noting that multiple-choice grammar tests often possess high face validity among students accustomed to traditional testing formats, yet such instruments may demonstrate weak construct validity for measuring authentic writing ability. Therefore, while face validity contributes to practical feasibility and stakeholder acceptance, it cannot substitute for rigorous validation procedures that establish whether assessments genuinely measure what they purport to assess (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Contemporary approaches to validation acknowledge face validity as one component within a comprehensive validity argument rather than treating it as a standalone criterion. Hubley and Zumbo (2011) suggest that soliciting stakeholder perceptions about assessment appropriateness can identify potential sources of construct-irrelevant variance—factors unrelated to the target construct that nevertheless influence scores. When test-takers view certain task features as unfair or irrelevant, their responses may reflect reactions to perceived invalidity rather than demonstrating actual writing competence. Consequently, attending to face validity during test development, while simultaneously pursuing empirical evidence through multiple validation strategies, represents a balanced approach that enhances both the technical quality and practical viability of writing assessments.

2.4.1.5 Consequential Validity

Consequential validity, a concept most thoroughly articulated by Messick (1989, 1995), extends traditional notions of validity beyond technical measurement properties to encompass the social consequences and ethical implications of test use. Messick argued that validity represents a unified construct integrating both the evidential basis for score interpretation and the consequential basis for score use, thereby making assessment developers and users accountable for both intended and unintended outcomes. This expanded conceptualization recognizes that tests function not merely as neutral measurement instruments but as social interventions that shape

educational opportunities, influence instructional practices, and affect individual life trajectories (Shepard, 1997).

In writing assessment contexts, consequential validity directs attention to how evaluation practices influence teaching and learning processes. Positive consequences might include enhanced instructional focus on important writing skills, increased student motivation to develop communication competencies, and more equitable identification of learners requiring additional support (Popham, 1997). For instance, when a writing assessment emphasizes authentic composing tasks and provides detailed diagnostic feedback, it may encourage instructors to adopt process-oriented pedagogies that support student development rather than focusing narrowly on test preparation (McNamara & Ryan, 2011). Such alignment between assessment and valued educational outcomes exemplifies what researchers' term positive washback or beneficial systemic impact (Wall, 2005).

However, assessments can also generate detrimental consequences that undermine educational quality and equity. High-stakes writing tests may narrow curriculum, prompting teachers to emphasize formulaic essay structures and test-taking strategies at the expense of authentic literacy development (Hillocks, 2002). This phenomenon, often called "teaching to the test," can result in instruction that privileges surface features easily measured in timed examinations while neglecting complex rhetorical abilities, critical thinking, and disciplinary writing conventions that develop more gradually (Linn, 2000). Furthermore, when writing assessments employ cultural or linguistic assumptions that disadvantage particular student populations, they may perpetuate inequities by misrepresenting the abilities of English language learners, students from non-dominant cultural backgrounds, or learners with disabilities (Valdés, 2001).

Messick (1995) distinguished between two dimensions of consequential validity: value implications and social consequences. Value implications concern the appropriateness of the construct being assessed—whether writing ability, as defined and measured by the test, reflects educationally important competencies rather than privileging arbitrary conventions or irrelevant skills. Social consequences encompass the actual effects of test use, including impacts on curriculum, instruction, student learning, and educational equity. Importantly, Messick emphasized that negative

consequences alone do not invalidate a test; rather, such outcomes raise questions about whether assessments should be used for particular purposes or whether mitigating measures are necessary to prevent harm (Shepard, 1997).

Establishing consequential validity requires systematic investigation of assessment impacts through multiple research approaches. These might include examining curricular changes following test implementation, documenting shifts in instructional practices, analyzing score distributions across demographic groups, and soliciting stakeholder perspectives on assessment effects (Lane & Stone, 2006). For writing assessment specifically, researchers might investigate whether portfolio-based evaluation encourages revision and multiple drafts, whether rubric criteria influence what teachers emphasize during writing instruction, or whether assessment formats differentially affect performance across cultural or linguistic groups (Huot, 2002).

Critics have debated the boundaries of consequential validity, with some arguing that test developers cannot be held responsible for all possible uses and misuses of assessment results (Popham, 1997). Nonetheless, contemporary validation standards explicitly recognize that consequences constitute an essential consideration in determining whether particular interpretations and uses of test scores are justified (American Educational Research Association et al., 2014). In the domain of writing assessment, where evaluation practices profoundly influence literacy instruction and can significantly impact student opportunities, attending to consequential validity represents both an ethical imperative and a practical necessity for ensuring that assessments serve their intended educational purposes while minimizing potential harms.

2.4.1.6 Ecological Validity

Ecological validity addresses the extent to which assessment conditions, tasks, and performance demands approximate the authentic contexts in which the target competency naturally occurs or will be applied (Bronfenbrenner, 1977; Schmuckler, 2001). Originally developed within psychological research to critique laboratory studies that failed to generalize beyond artificial experimental settings, the concept has gained prominence in educational assessment as scholars increasingly recognize that decontextualized testing environments may produce results bearing limited

relationship to real-world performance (Choi, Lee, & Kang, 2009). For writing assessment, ecological validity concerns whether evaluation tasks, administration conditions, and scoring criteria reflect the actual writing situations learners encounter in academic, professional, or civic life (Lewkowicz, 2000).

Traditional writing tests often compromise ecological validity by imposing constraints rarely present in authentic composing contexts. Timed essay examinations, for instance, require students to generate, organize, and produce text within compressed timeframes, typically without access to reference materials, revision opportunities, or authentic audiences—conditions starkly different from most consequential writing situations (Huot, 2002). Similarly, when assessment prompts dictate artificial topics divorced from students' actual disciplinary studies or personal interests, the resulting performances may inadequately represent their capabilities to write purposefully about subjects they genuinely understand and care about (Hamp-Lyons & Condon, 2000). Such ecological invalidity raises questions about whether scores derived from these assessments support valid inferences about students' abilities to succeed in the writing contexts that matter educationally or professionally.

Several factors contribute to ecological validity in writing assessment. Task authenticity involves designing prompts and activities that resemble genuine rhetorical situations, complete with realistic purposes, audiences, and genres (Bachman & Palmer, 1996). For example, asking business communication students to compose an actual letter addressing a workplace scenario demonstrates higher ecological validity than requesting a generic essay on communication principles. Contextual fidelity concerns whether assessment conditions replicate relevant features of target performance environments, including access to resources, time constraints, and collaborative opportunities (McNamara, 1996). A writing assessment for academic purposes might enhance ecological validity by permitting dictionary use, allowing sufficient time for revision, and accepting varied discourse structures rather than mandating a single organizational template.

Process authenticity represents another dimension of ecological validity, addressing whether assessments capture the recursive, iterative nature of skilled writing. Portfolios that include multiple drafts, reflective commentary, and evidence of

revision processes potentially offer greater ecological validity than single-draft timed tests because they acknowledge that proficient writers typically engage in planning, drafting, feedback-seeking, and extensive revision (Elbow & Belanoff, 1997). Similarly, performance assessments that embed writing within broader project-based activities—such as conducting research and composing reports—may better approximate authentic literacy practices than isolated composition tasks (Wiggins, 1998).

Research suggests that ecological validity influences not only the generalizability of assessment results but also the nature of performances themselves. When students perceive tasks as authentic and meaningful, they often demonstrate higher levels of engagement, invest greater cognitive effort, and produce qualitatively different writing than when completing obviously artificial assignments (Frederiksen & Collins, 1989). This phenomenon implies that ecologically valid assessments may actually elicit more representative samples of student competence, thereby enhancing construct validity alongside practical applicability. Conversely, ecologically invalid tasks may underestimate abilities by creating artificial barriers or overestimate them by simplifying demands beyond real-world requirements (Messick, 1994).

However, pursuing maximum ecological validity involves inherent trade-offs with other assessment priorities. Highly contextualized, authentic tasks may reduce standardization, complicate scoring reliability, and create inequities if students possess differential familiarity with particular contexts (Kane, 2006). For instance, a writing assessment embedded within a specific disciplinary context might advantage students with relevant background knowledge while disadvantaging those encountering the field for the first time, potentially confounding writing ability with content knowledge. Additionally, some authentic writing situations involve extended timelines, collaborative authorship, or extensive resource consultation—conditions difficult to replicate within practical assessment constraints (Cumming, 2002).

Effective assessment design requires balancing ecological validity against these competing considerations, seeking optimal correspondence with authentic performance contexts while maintaining feasibility, fairness, and psychometric soundness. Cumming (2002) suggests that rather than attempting perfect ecological

simulation, writing assessments should capture the essential cognitive processes, rhetorical challenges, and linguistic demands characteristic of target writing situations. This approach acknowledges that all assessments necessarily involve some abstraction from authentic contexts while striving to preserve those features most critical for valid inference about real-world writing capabilities. By systematically analyzing the writing demands students will face beyond the assessment context and designing tasks that preserve key elements of those situations, educators can enhance ecological validity and thereby strengthen the meaningfulness and utility of writing assessment results.

2.4.2 Key Aspects of Validity

While validity types provide frameworks for examining different evidence sources, several fundamental principles undergird all validity considerations in writing assessment. These key aspects represent essential qualities that valid assessments must demonstrate, cutting across specific validity categories to address fundamental questions about measurement quality, interpretive soundness, and ethical assessment practice.

2.4.2.1 Accuracy in Measurement

Accuracy in measurement constitutes a foundational aspect of validity, addressing the fundamental question of whether assessment results faithfully represent the writing competencies they purport to evaluate (Cronbach & Meehl, 1955). This dimension concerns the technical precision with which tests capture the target construct while minimizing errors that distort score meaning. In writing assessment contexts, accuracy requires that evaluation procedures systematically differentiate among examinees based on genuine variations in writing proficiency rather than reflecting construct-irrelevant factors such as rater bias, prompt difficulty variations, or scoring inconsistencies (Weigle, 2002).

Several potential sources of measurement error threaten accuracy in writing assessment. Rater variability represents a prominent concern, as research consistently documents that different evaluators may assign disparate scores to

identical writing samples due to divergent standards, preferences, or interpretive frameworks (Lumley, 2005). Even individual raters demonstrate inconsistency over time, with factors like fatigue, mood, or sequence effects influencing judgments. Systematic rater bias—such as consistently favoring particular rhetorical styles, penalizing non-native language features, or holding differential expectations for students from various demographic groups—introduces construct-irrelevant variance that compromises measurement accuracy (Engelhard, 1994).

Task and prompt characteristics similarly affect accuracy. When assessment prompts vary substantially in difficulty, familiarity, or the writing abilities they elicit, scores may reflect which particular task students encountered rather than their underlying writing competence (Huot & Neal, 2006). A student who performs well on a narrative prompt might struggle with argumentative writing, raising questions about whether a single task provides an accurate gauge of general writing ability. Furthermore, cultural bias in prompt content—such as topics requiring specific background knowledge or reflecting particular worldviews—can systematically advantage or disadvantage certain student groups, thereby reducing measurement accuracy for diverse populations (Fox, 2004).

Establishing measurement accuracy requires multiple strategies. Rigorous rater training that emphasizes scoring criteria, provides calibration exercises, and monitors consistency helps minimize evaluator-related errors (Weigle, 1998). Using multiple raters and aggregating their independent judgments can reduce individual biases through statistical averaging. Developing carefully piloted prompts that have been tested with diverse student samples and revised to minimize construct-irrelevant difficulty or cultural bias enhances task-related accuracy. Additionally, assessing writing through multiple tasks or occasions reduces the influence of any single prompt's peculiarities, providing a more stable and accurate estimate of overall competence (Brennan, 2001).

Measurement accuracy also intersects with scoring methodology. Analytic rubrics that separately evaluate distinct dimensions of writing may enhance accuracy by directing rater attention to specific features and reducing holistic impressions' susceptibility to halo effects (Barkaoui, 2010). However, overly atomized scoring schemes risk losing sight of writing's integrated nature, potentially compromising

accuracy by failing to capture how effectively various elements combine to achieve rhetorical purposes. Balancing specificity with recognition of writing's holistic character represents an ongoing challenge in pursuing measurement accuracy (Broad, 2003).

Ultimately, accuracy in measurement serves as a necessary but insufficient condition for validity. Even highly precise scores that consistently differentiate among students may prove invalid if they measure the wrong construct or support inappropriate inferences. Nonetheless, without reasonable measurement accuracy, subsequent validity considerations become moot, as unreliable data cannot sustain meaningful interpretations regardless of other assessment strengths (American Educational Research Association et al., 2014).

2.4.2.2 Meaningfulness of Inferences

Meaningfulness of inferences addresses whether the interpretations drawn from assessment results constitute warranted, reasonable, and useful conclusions about student writing capabilities (Kane, 2013). This aspect recognizes that tests themselves do not possess validity; rather, validity attaches to the inferences, interpretations, and uses that follow from test scores. An assessment might yield highly reliable scores demonstrating strong technical properties, yet if the conclusions drawn from those scores misrepresent student abilities or lead to inappropriate decisions, validity remains compromised (Messick, 1989).

In writing assessment, meaningfulness concerns whether scores support the specific claims stakeholders intend to make. When instructors use assessment results to conclude that a student "demonstrates proficient academic writing ability," this inference implies a range of capabilities including coherent organization, appropriate register, effective argument development, and linguistic control sufficient for disciplinary communication. For such an inference to be meaningful, assessment tasks must adequately sample these diverse competencies, and scoring procedures must weight them appropriately (Cumming, Kantor, & Powers, 2002). If the assessment actually emphasizes grammatical accuracy at the expense of rhetorical effectiveness, inferences about overall academic writing proficiency lack warrant—the evidence does not support the interpretation.

Establishing meaningful inferences requires articulating explicit validity arguments that specify intended interpretations, identify supporting evidence, and acknowledge alternative explanations (Kane, 2006, 2013). This argument-based approach to validation treats validity as a logical process of marshaling evidence and reasoning to support or refute proposed interpretations. For a writing placement test, the validity argument might specify that scores will be interpreted to mean students possess the writing skills necessary to succeed in particular courses. Supporting this interpretation requires evidence that test performance predicts academic writing success, that assessment tasks reflect actual course demands, and that placement decisions based on scores lead to appropriate instructional placements (Huot, 1996).

The meaningfulness of inferences can be undermined by various factors. Construct underrepresentation occurs when assessments fail to capture important dimensions of the target construct, leading to overly narrow inferences that mischaracterize student abilities (Messick, 1995). A writing test comprising only grammar and usage items, for instance, provides inadequate basis for inferences about composing ability more broadly. Conversely, construct-irrelevant variance introduces extraneous factors that inflate or deflate scores without reflecting actual writing proficiency, such as handwriting legibility in holistically scored essays or reading comprehension demands in prompt-dependent tasks (Haladyna & Downing, 2004).

Context also shapes inference meaningfulness. Inferences appropriate in one situation may prove unwarranted in another, even when based on identical assessment procedures. For example, using diagnostic writing assessments to make high-stakes admissions decisions stretches inference beyond original purposes, potentially invalidating interpretations designed for formative instructional guidance (Kane, 2013). Similarly, inferences from assessments developed for native speakers may lack meaningfulness when applied to English language learners, whose performances reflect different combinations of linguistic development, rhetorical knowledge, and communicative competence (Cumming, 1997).

Validation of inference meaningfulness thus requires continuously examining whether interpretations align with evidence, considering alternative explanations for observed performances, and recognizing the boundaries within which particular inferences remain warranted. This ongoing process acknowledges that validity is not

an all-or-nothing property assessment either possess or lack, but rather a matter of degree depending on proposed interpretations, available evidence, and intended uses (Cronbach, 1988).

2.4.2.3 Fairness

Fairness represents a fundamental ethical dimension of validity, addressing whether assessments provide all examinees equitable opportunities to demonstrate their writing abilities without systematic advantage or disadvantage based on characteristics irrelevant to the construct being measured (American Educational Research Association et al., 2014). In increasingly diverse educational contexts, fairness concerns have assumed heightened prominence as researchers document how seemingly neutral assessment practices can perpetuate inequities by privileging certain linguistic backgrounds, cultural experiences, or educational opportunities while marginalizing others (Kunnan, 2000).

Multiple perspectives inform fairness considerations in writing assessment. Procedural fairness concerns whether all students receive comparable administration conditions, including adequate time, appropriate accommodations for disabilities, clear instructions, and unbiased rating procedures (Xi, 2010). When some examinees encounter noisier testing environments, receive inadequate directions, or have their work evaluated by harsher raters, procedural inequities compromise fairness regardless of assessment content quality. Ensuring procedural fairness requires standardized protocols, systematic monitoring of administration conditions, and provision of necessary accommodations without stigmatization (Sireci, Scarpati, & Li, 2005).

Content fairness addresses whether assessment tasks, prompts, and required knowledge distribute opportunities equitably across diverse student populations. Topics presuming specific cultural knowledge, language varieties, or life experiences may systematically disadvantage examinees lacking that background (Solórzano, 2008). For instance, prompts requiring students to discuss holiday traditions, family structures, or recreational activities common in dominant cultures may advantage native-born students while creating additional cognitive burdens for immigrants or international learners who must navigate cultural translation alongside writing

demands. Similarly, decontextualized academic topics may favor students from educationally advantaged backgrounds with broader exposure to such discourse (Valdés, Bunch, Snow, & Lee, 2005).

Linguistic fairness poses particular challenges in EFL writing assessment, where students' developing English proficiency intersects with writing ability in complex ways. Distinguishing language development from writing competence proves difficult when assessment necessarily occurs through language (Weigle, 2002). Should evaluations penalize organizational weaknesses resulting from limited familiarity with English rhetorical conventions? Do spelling errors carry the same significance for language learners as for native speakers? Failure to account for such distinctions may systematically underestimate the writing abilities of EFL students, invalidating inferences about their compositional competence apart from language proficiency (Crusan, 2010).

Fairness also encompasses consequential dimensions, examining whether assessment results differentially impact various student groups in ways unrelated to actual ability differences. When writing assessments systematically produce lower scores for particular demographic groups, and when those scores determine access to educational opportunities, placement decisions, or certification, potential inequities warrant careful investigation (Zwick, 2006). Disparate impact alone does not necessarily indicate unfairness—genuine ability differences might exist. However, when score disparities cannot be explained by construct-relevant factors, or when assessments measure abilities peripheral to genuine writing competence, fairness questions arise (Willingham & Cole, 1997).

Enhancing fairness requires proactive design decisions and ongoing monitoring. Universal design principles advocate creating assessments accessible to diverse learners from the outset rather than retrofitting accommodations (Thompson, Johnstone, & Thurlow, 2002). This might involve offering prompt choices, permitting dictionary use for language learners, or designing rubrics that appropriately weight rhetorical effectiveness versus surface-level linguistic features. Systematic review of assessment materials by diverse stakeholders can identify potentially biased content before implementation (Solano-Flores, 2008). Post-administration analyses examining score distributions, item functioning, and rater

patterns across demographic groups help detect fairness problems requiring correction (Zieky, 2006).

Importantly, fairness does not require identical outcomes across groups or elimination of all performance differences. Rather, it demands that assessments measure the intended construct as accurately and equitably as possible for all examinees, that score differences reflect genuine variations in the target ability rather than construct-irrelevant factors, and that assessment uses respect principles of justice and avoid perpetuating systemic inequities (Dorans & Cook, 2016). In writing assessment particularly, where language, culture, and power intersect complexly, fairness represents an ongoing commitment to critically examining whose writing abilities are recognized and valued, whose communicative strengths might be overlooked, and how assessment practices can promote rather than impede equitable educational opportunities.

2.4.3 Factors Influencing Validity

The validity of writing assessments emerges not from any single design decision but from the complex interaction of multiple factors that collectively determine whether evaluations support warranted inferences about student competence. Understanding these influential factors enables assessment developers and users to make informed choices that enhance validity while recognizing the inherent tensions and trade-offs involved in evaluation design. Three particularly consequential factors—rubric design, scoring method, and number of assessment occasions—merit detailed examination given their substantial impact on the quality and defensibility of writing assessment practices.

2.4.3.1 Rubric Design

The design of assessment rubrics exerts profound influence on validity by shaping what evaluators attend to, how performances are interpreted, and ultimately what inferences can be justified from resulting scores. Rubrics function as operational definitions of the writing construct, translating abstract notions of competence into concrete criteria and performance descriptors that guide judgment (Popham, 2008). Consequently, rubric design decisions directly affect construct representation—the extent to which assessments capture the full range of knowledge, skills, and abilities

constituting proficient writing (Messick, 1995).

Well-designed rubrics enhance validity by providing explicit, comprehensive articulation of valued writing qualities. When rubric criteria systematically address essential dimensions of writing competence—such as rhetorical effectiveness, organizational coherence, idea development, audience awareness, and linguistic control—they help ensure that evaluations reflect the construct's complexity rather than focusing narrowly on easily observable surface features (Broad, 2003). For instance, a rubric emphasizing only grammatical accuracy and mechanical correctness would demonstrate construct underrepresentation, failing to capture critical aspects of writing ability like purposeful communication, critical thinking, or genre awareness. Such narrow operationalization undermines content validity by misaligning assessment with comprehensive learning objectives (Moskal & Leydens, 2000).

The specificity and clarity of rubric descriptors similarly influence validity. Vague language like "demonstrates good organization" or "uses appropriate vocabulary" leaves substantial room for subjective interpretation, potentially reducing inter-rater reliability and introducing construct-irrelevant variance as different evaluators operationalize these terms idiosyncratically (Jonsson & Svingby, 2007). Conversely, overly prescriptive rubrics that specify rigid structural formulas or mandate particular stylistic choices may artificially constrain writing and privilege formulaic responses over authentic, rhetorically sophisticated performances (Wilson, 2006). Effective rubric design thus requires balancing specificity sufficient to guide consistent evaluation with flexibility that acknowledges legitimate variation in how competent writers achieve rhetorical goals.

The number and nature of criteria included in rubrics present another validity consideration. Analytic rubrics with multiple distinct dimensions provide rich diagnostic information and can enhance content validity by ensuring attention to diverse competencies (Brookhart, 2013). However, excessive atomization—breaking writing into numerous separate components—risks losing sight of the integrated, holistic nature of effective communication. Research suggests that writers do not develop discrete skills in isolation; rather, various competencies interact dynamically during composing (Cumming, Kantor, & Powers, 2002). Rubrics that

treat these elements as entirely independent may therefore misrepresent the construct's essential nature.

Furthermore, rubric design choices carry consequential validity implications by influencing instruction and learning. Rubrics emphasizing easily teachable and assessable features may prompt teachers to focus disproportionately on those elements, potentially narrowing curriculum (Kohn, 2006). If rubrics heavily weight surface correctness while minimizing rhetorical effectiveness or critical engagement with ideas, instruction may shift toward error avoidance rather than meaningful communication. Conversely, rubrics that articulate sophisticated standards for argumentation, evidence use, and audience adaptation can encourage instructional practices supporting these higher-order competencies (Andrade, 2000).

Cultural and linguistic assumptions embedded in rubrics also affect validity, particularly for diverse student populations. Criteria presuming familiarity with specific rhetorical conventions, discourse patterns, or stylistic preferences reflecting dominant cultural traditions may systematically disadvantage writers from other backgrounds (Crusan, 2010). For EFL contexts, rubrics must thoughtfully address the relationship between language development and writing ability, establishing appropriate expectations that distinguish linguistic proficiency from compositional competence while recognizing their interdependence (Weigle, 2002). Rubric design that fails to account for legitimate linguistic variation or cultural differences in rhetorical approaches compromises fairness and construct validity.

2.4.3.2 Scoring Method

The method employed to transform observations of writing performances into numerical or categorical scores significantly influences validity by affecting what information is captured, how consistently judgments are rendered, and what interpretations scores can support. Different scoring approaches—holistic, analytic, primary trait, or portfolio-based—embody distinct philosophical assumptions about writing's nature and assessment's purposes, yielding scores with varying validity for different inferences (White, 1994).

Holistic scoring, which assigns a single overall rating based on general impression of quality, offers efficiency and acknowledges writing's integrated character (Weigle,

2002). This method aligns with views of writing as a unified communicative act where various elements combine synergistically to achieve rhetorical effects that cannot be reduced to component parts. Holistic scores may demonstrate adequate validity for broad placement decisions or general proficiency judgments when raters have been thoroughly trained to internalize comprehensive scoring criteria (White, 1985). However, the method's limited diagnostic specificity constrains validity for formative purposes, as single scores provide insufficient information about particular strengths or weaknesses requiring instructional attention (Hamp-Lyons, 2003). Additionally, holistic scoring may be susceptible to halo effects where prominent features—positive or negative—disproportionately influence overall ratings, potentially introducing construct-irrelevant variance (Lumley, 2005).

Analytic scoring addresses some holistic methods' limitations by evaluating writing across multiple distinct dimensions, yielding separate scores for criteria like content, organization, language use, and mechanics (Jacobs et al., 1981). This approach enhances content validity by ensuring systematic attention to diverse competencies and provides detailed diagnostic information supporting instructional decision-making (Barkaoui, 2010). Research indicates that analytic scoring can improve inter-rater reliability by directing evaluator attention to specific features and reducing the influence of overall impressions (Jonsson & Svingby, 2007). However, analytic methods require substantially more rating time and may artificially fragment writing into independent components, potentially compromising construct validity by failing to capture how elements interact to create effective communication (Broad, 2003). Furthermore, when analytic rubrics weight dimensions equally despite their varying importance for particular writing purposes, resulting scores may misrepresent overall competence.

Primary trait scoring focuses specifically on the rhetorical effectiveness of writing in relation to particular communicative purposes, assessing how well writers achieve specific objectives given particular audiences and situations (Lloyd-Jones, 1977). This method can enhance construct validity for assessments targeting defined rhetorical competencies by concentrating evaluation on features directly relevant to communicative success rather than generic writing quality. Primary trait approaches align well with authentic assessment philosophies emphasizing context-specific performance (Wiggins, 1998). However, the method's narrow focus may limit

generalizability of inferences about broader writing ability, as strong performance on one rhetorical task does not necessarily predict success across diverse writing situations (Huot, 2002).

The involvement of multiple raters and procedures for synthesizing their judgments constitute another scoring methodology dimension affecting validity. Single-rater designs prove economical but introduce substantial individual variance that may reflect idiosyncratic preferences rather than systematic quality differences (Lumley, 2002). Multiple independent ratings that are subsequently averaged or adjudicated can enhance reliability and reduce bias, strengthening validity (Weigle, 1998). However, consensus scoring where raters discuss performances before assigning scores may compromise independence and introduce group dynamics that either increase consistency through calibration or introduce new sources of variance through social pressure (Lumley, 2005).

Scoring method choices also carry consequential validity implications. Methods providing detailed analytical feedback can support learning by helping students understand specific areas for improvement, potentially enhancing instructional effectiveness (Panadero & Jonsson, 2013). Conversely, scoring approaches that obscure the relationship between performances and scores may diminish assessment's educational value and reduce student engagement with feedback. Additionally, when high-stakes decisions rest on scores from methods with recognized limitations—such as using holistic ratings for detailed diagnostic purposes—validity concerns arise regarding appropriateness of inferences and uses.

2.4.3.3 Number of Tests

The number of writing tasks or assessment occasions included in evaluation systems substantially influences validity by affecting score generalizability, construct representation, and the reliability of inferences drawn from results. Single-occasion, single-task assessments—while administratively convenient—demonstrate notable limitations that compromise validity for most important educational purposes (Huot & Neal, 2006).

Research consistently documents considerable task-specific variation in writing performance, with individuals often demonstrating substantially different

competence levels across prompts, genres, or rhetorical situations (Cumming *et al*., 2002). This variability reflects multiple factors including differential familiarity with topics, varying strengths across discourse modes, and the influence of prompt characteristics on the particular abilities elicited. When assessment relies on a single task, scores may reflect the match or mismatch between that specific prompt and individual students' knowledge, interests, and competencies rather than representing stable writing ability (Dunbar *et al*., 1991). Such task specificity introduces construct-irrelevant variance that undermines validity by making scores depend partly on which particular prompt students encountered.

Increasing the number of assessment tasks enhances validity by reducing measurement error associated with task specificity. Generalizability theory provides a framework for understanding how different sources of variation—including persons, tasks, occasions, and raters—contribute to score variability (Brennan, 2001). Studies employing generalizability analyses consistently demonstrate that multiple tasks yield more dependable estimates of writing competence than single performances, with reliability coefficients improving substantially when assessments include two or three distinct writing samples (Shavelson & Webb, 1991). This improved reliability strengthens validity by ensuring that inferences rest on more stable, representative evidence of student capabilities.

Multiple tasks also enhance content validity by enabling broader sampling of the writing construct. Different prompts can elicit varied discourse modes—narrative, expository, argumentative, analytical—thereby assessing students' versatility across rhetorical purposes (Hamp-Lyons & Condon, 2000). Similarly, diverse task types might include timed impromptu writing, extended process-based compositions, and research-integrated projects, collectively capturing the multifaceted nature of writing competence more comprehensively than any single format. This broader sampling reduces construct underrepresentation and supports richer, more warranted inferences about students' overall writing abilities across contexts.

Portfolio assessment represents an extended application of multiple-task principles, typically incorporating numerous writing samples collected over time (Elbow & Belanoff, 1997). Portfolios offer substantial validity advantages by documenting developmental trajectories, including evidence of revision and improvement, and

allowing students to demonstrate competence across diverse authentic tasks. The longitudinal perspective portfolios provide supports inferences about students' learning growth and sustained capabilities rather than momentary performances potentially influenced by transient factors (Hamp-Lyons & Condon, 2000). However, portfolios present validity challenges including standardization difficulties, questions about work authenticity when composition occurs outside supervised settings, and substantial time demands for comprehensive evaluation (Huot, 2002).

The timing and spacing of multiple assessment occasions also influence validity. Assessments administered in close temporal proximity may demonstrate task-to-task correlation partly reflecting transient factors like student mood, energy level, or recent instruction rather than stable competence (Dunbar *et al*., 1991). Conversely, assessments distributed across extended periods can capture genuine development and provide more representative sampling of typical performance. However, when lengthy intervals separate assessments, intervening instruction and maturation may complicate interpretation of score patterns.

Practical constraints inevitably limit the number of tasks feasible in most assessment contexts. Each additional writing sample requires student time for composition and evaluator time for scoring, creating resource demands that can become prohibitive (Weigle, 2002). Assessment designers must therefore balance validity benefits of multiple tasks against practical limitations, seeking optimal sampling strategies that maximize information while remaining implementable. Research suggests that even modest increases from one to two or three tasks substantially enhance generalizability, with diminishing returns for additional samples (Breland, 1983). Strategic task selection—ensuring diversity in prompts, topics, and rhetorical demands—can maximize the validity gains achievable with limited numbers of assessments.

2.4.4 Ensuring Validity in Writing Assessment

While understanding factors that influence validity provides necessary foundation, translating this knowledge into practice requires concrete strategies for designing and implementing assessments that support warranted inferences. Four particularly

important approaches—systematic rubric use, consistency checks, comprehensive attention to writing dimensions, and authentic task design—offer practical means of enhancing validity in writing assessment contexts.

2.4.4.1 Use of Rubrics

The systematic, principled use of well-designed rubrics constitutes a fundamental strategy for enhancing validity in writing assessment. As discussed previously, rubrics operationalize the writing construct by articulating evaluation criteria and performance standards, thereby providing explicit frameworks that guide both instruction and assessment (Brookhart, 2013). However, merely possessing a rubric proves insufficient; validity depends critically on how rubrics are developed, implemented, and utilized throughout the assessment process.

Effective rubric development begins with careful construct definition grounded in clear articulation of learning objectives and authentic writing demands students will encounter (Moskal, 2000). This process should involve multiple stakeholders including experienced writing instructors, assessment specialists, and when appropriate, students themselves, ensuring that criteria reflect shared understanding of valued competencies rather than individual preferences (Broad, 2003). Systematic construct analysis—examining what knowledge, skills, and abilities constitute proficient writing for particular purposes and contexts—helps identify essential dimensions requiring assessment while avoiding construct irrelevance (Messick, 1995).

Rubric criteria should comprehensively represent the target construct while maintaining practical feasibility for implementation. This requires thoughtful decisions about which writing dimensions to include, how to weight them relative to one another, and what level of specificity to employ in performance descriptors (Popham, 2008). For EFL writing assessment specifically, rubrics must thoughtfully address relationships between linguistic development and compositional ability, establishing appropriate standards that neither conflate these dimensions nor treat them as entirely separate (Weigle, 2002). Clear descriptors should characterize qualitatively distinct performance levels using concrete, observable language that helps raters recognize salient differences while avoiding artificial precision that

suggests greater measurement accuracy than assessment methods can support (Sadler, 2009).

Pilot testing rubrics with actual student work samples proves essential for validation. This process involves applying draft rubrics to diverse performances, examining whether criteria enable meaningful differentiation across quality levels, identifying ambiguities or gaps in descriptors, and verifying that the full scoring scale functions appropriately (Tierney & Simon, 2004). Pilot testing should intentionally include work from varied student populations to ensure rubrics function equitably and do not systematically advantage or disadvantage particular groups (Solano-Flores, 2008). Iterative refinement based on pilot results strengthens content validity and practical utility before operational implementation.

Thorough rater training in rubric use represents another critical element. Research demonstrates that even well-designed rubrics cannot overcome inadequate rater preparation; evaluators must develop shared understanding of criteria, internalize standards represented by performance descriptors, and practice applying rubrics consistently (Weigle, 1998). Effective training programs include multiple components: reviewing scoring criteria and their rationale, examining anchor papers representing different performance levels with discussion of distinguishing features, engaging in practice scoring with feedback, and establishing acceptable agreement thresholds before independent rating begins (Lumley, 2005). Periodic recalibration sessions help maintain consistency over extended scoring periods or across multiple rating occasions.

Ongoing monitoring of rubric functioning during operational use supports continued validity. This includes tracking inter-rater reliability to ensure consistent application, examining score distributions to identify unexpected patterns suggesting rubric problems, and soliciting rater feedback about ambiguities or implementation challenges (Engelhard, 1994). Such monitoring may reveal that certain criteria prove difficult to apply reliably, that some performance descriptors require clarification, or that the rubric fails to differentiate adequately at particular score levels. Responsive revision based on implementation evidence demonstrates commitment to validity as an ongoing process rather than a one-time achievement (Kane, 2013).

Finally, rubrics should be shared transparently with students, ideally before they begin writing tasks. This practice enhances face validity by clarifying expectations, supports learning by helping students understand quality standards, and promotes fairness by ensuring all examinees access the same information about evaluation criteria (Andrade, 2000). Student involvement in rubric development or modification—examining work samples, discussing quality indicators, and articulating standards—can deepen understanding while increasing perceived legitimacy (Stiggins, 2001). Such transparency aligns assessment more closely with instruction and helps ensure that evaluation serves learning rather than merely measuring it.

2.4.4.2 Consistency Checks

Establishing and maintaining consistency in scoring represents an essential validity strategy, as unreliable evaluations cannot support warranted inferences regardless of other assessment strengths. Inconsistency introduces measurement error that obscures true differences in writing competence, potentially leading to invalid conclusions about student abilities and inappropriate educational decisions (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Multiple forms of consistency checks—examining agreement across raters, occasions, and tasks—help identify and minimize threats to reliable, valid assessment.

Inter-rater reliability checks constitute the most common consistency monitoring approach, examining the extent to which independent evaluators assign similar scores to identical writing samples. Research consistently documents substantial rater variability in writing assessment, with different evaluators sometimes reaching markedly discrepant judgments even when applying the same rubric (Lumley, 2002). This variability may reflect genuine disagreement about quality, differential weighting of criteria, idiosyncratic preferences, or inconsistent application of scoring standards. Whatever its source, excessive rater disagreement undermines validity by introducing construct-irrelevant variance—score differences reflecting evaluator characteristics rather than performance quality.

Multiple statistical indices can quantify inter-rater reliability, each with particular advantages and interpretive considerations. Percentage agreement provides intuitive

understanding but fails to account for chance agreement and treats all disagreements equally regardless of magnitude (Stemler, 2004). Correlation coefficients capture relative consistency—whether raters rank-order performances similarly—but can appear acceptable even when raters systematically apply different standards that shift all scores uniformly. Kappa coefficients adjust for chance agreement and prove appropriate for categorical ratings, while intraclass correlations suit continuous scores and distinguish different sources of variance (Hallgren, 2012). Generalizability theory offers particularly sophisticated consistency analysis, partitioning variance attributable to persons, raters, tasks, and their interactions, thereby illuminating specific sources of inconsistency requiring attention (Brennan, 2001).

Acceptable reliability thresholds depend on assessment purposes and stakes. High-stakes decisions affecting students' educational opportunities demand strong inter-rater reliability, typically with correlations or kappa values exceeding .80 or .90 (Nunnally & Bernstein, 1994). Lower-stakes classroom assessments might tolerate somewhat greater variability, though even formative evaluation requires sufficient consistency to provide meaningful feedback. When inter-rater reliability falls short, possible responses include enhanced rater training, rubric clarification, increased number of raters with score averaging, or adoption of consensus scoring procedures where disagreements trigger discussion and resolution (Lumley, 2005).

Intra-rater reliability—individual evaluators' consistency over time—represents another important dimension often receiving insufficient attention. Factors including fatigue, order effects, and changing internalized standards can cause single raters to judge identical performances differently on separate occasions (Saal, Downey, & Lahey, 1980). Including repeated samples—where some papers appear multiple times throughout a scoring session—enables detection of individual drift and provides evidence about temporal stability. Finding that raters assign markedly different scores to the same performance at different times signals problems requiring intervention, such as frequent recalibration breaks or reduced scoring session length.

Cross-task consistency examines whether students demonstrate similar performance levels across different writing prompts or occasions. While perfect consistency

cannot be expected given legitimate task specificity, extreme variation may indicate assessment problems such as prompts eliciting very different abilities or rubrics functioning inconsistently across task types (Dunbar et al., 1991). Correlation analyses examining score relationships across multiple tasks, or generalizability studies decomposing person-by-task interactions, help evaluate whether assessments provide stable estimates of writing competence or reflect primarily task-specific factors (Shavelson & Webb, 1991). Low cross-task consistency suggests need for increased number of assessment occasions or more careful prompt development to ensure comparable difficulty and construct relevance.

Implementing consistency checks requires systematic data collection and analysis infrastructure. For large-scale assessments, this might involve double-scoring samples of work with subsequent disagreement analysis, establishing ongoing quality control procedures, and employing specialized software tracking rater performance patterns (Bejar, 2012). Classroom assessments can incorporate simpler consistency procedures such as periodic cross-checking where teachers exchange and score subsets of each other's students' papers, or self-monitoring where instructors re-score samples from early scoring sessions to detect drift. Regardless of scale, treating consistency monitoring as integral to assessment rather than optional adds substantial validity evidence.

2.4.4.3 Focus on All Aspects of Writing

Valid writing assessment requires comprehensive attention to the multiple dimensions constituting writing competence rather than narrow focus on easily measured surface features. The tendency to emphasize grammatical accuracy, mechanical correctness, and other readily quantifiable elements while minimizing rhetorical effectiveness, critical thinking, or genre awareness represents a persistent threat to construct validity (Broad, 2003). Ensuring that assessments and scoring procedures adequately attend to all valued aspects of writing constitutes an essential validity strategy.

Contemporary understandings recognize writing as complex, multifaceted activity involving numerous interrelated competencies. Hyland (2003) characterizes writing as simultaneously a cognitive process requiring planning and problem-solving, a

social practice shaped by context and community conventions, and a textual product demonstrating linguistic and rhetorical features. Effective writing demands content knowledge, rhetorical awareness, organizational skill, linguistic control, genre understanding, audience sensitivity, and critical thinking ability (Grabe & Kaplan, 1996). Assessments focusing predominantly on any single dimension—whether grammar, organization, or argumentation-demonstrate construct underrepresentation that undermines content validity.

The historical tendency to privilege surface correctness in writing evaluation reflects multiple factors including ease of identification, relative objectivity of mechanical errors, and deeply embedded beliefs about grammar's centrality to writing quality (Connors & Lunsford, 1988). However, research consistently demonstrates weak relationships between surface feature control and overall writing effectiveness; error-free prose can be rhetorically impoverished while texts with various mechanical issues may nonetheless communicate powerfully (Haswell, 1988). For EFL writers particularly, exclusively error-focused assessment risks systematically undervaluing authentic communicative competence while overemphasizing linguistic development that proceeds gradually (Ferris, 2003).

Comprehensive assessment requires that evaluation criteria and scoring procedures systematically address multiple writing dimensions with appropriate relative weighting. Analytic rubrics prove particularly useful for ensuring such breadth, with separate scales for distinct yet complementary aspects like content quality, organizational coherence, rhetorical effectiveness, source integration, linguistic accuracy, and mechanical correctness (Jacobs et al., 1981). However, simply including multiple criteria proves insufficient if raters, through training or personal disposition, attend disproportionately to particular dimensions while minimizing others. Explicit training emphasizing all rubric components, with practice identifying both strengths and weaknesses across different quality aspects, helps counteract tendencies toward narrow focus (Weigle, 1998).

Task design significantly influences which writing dimensions assessments ultimately measure. Prompts emphasizing personal narrative may foreground descriptive ability and experiential content while providing limited opportunity to demonstrate analytical reasoning or evidence-based argumentation (Hamp-Lyons,

1991). Research synthesis assignments necessarily involve source integration and critical evaluation skills that simpler tasks do not require. Genre-specific writing draws on knowledge of particular discourse conventions that cannot be assessed through generic essay prompts. Ensuring comprehensive construct coverage therefore requires thoughtful task selection or, more reliably, multiple diverse prompts that collectively elicit varied competencies (Cumming et al., 2002).

Formative assessment contexts particularly benefit from comprehensive attention to writing dimensions, as detailed diagnostic information across multiple aspects enables targeted instructional response. When feedback addresses not only surface errors but also higher-order concerns like argumentation quality, audience adaptation, and organizational effectiveness, students receive guidance supporting development across the full spectrum of writing competencies (Ferris, 2003). Conversely, feedback focusing narrowly on grammar may inadvertently communicate those other dimensions matter less, potentially misdirecting learning efforts.

Balancing attention across writing aspects requires acknowledging legitimate tensions. Cognitive load limitations constrain how much information raters can process simultaneously; attempting to evaluate too many dimensions concurrently may reduce attention to each (Lumley, 2005). Additionally, different writing purposes and contexts appropriately prioritize certain competencies. Technical documentation demands precision and clarity but may minimize stylistic creativity, while literary analysis privileges interpretive sophistication and textual engagement. Valid assessment therefore requires not uniform treatment of all writing dimensions across contexts but rather appropriate emphasis aligned with specific rhetorical situations and learning objectives (Hyland, 2003).

2.4.4.4 Authentic Tasks: Real-World Writing Tasks

Incorporating authentic tasks that meaningfully approximate real-world writing situations represents a powerful validity strategy, enhancing both construct representation and the generalizability of inferences from assessment to consequential performance contexts. Authenticity in assessment design addresses fundamental questions about whether evaluation tasks elicit the actual competencies

students need beyond testing situations and whether performances under assessment conditions predict capability in authentic contexts where writing matters (Wiggins, 1998).

Traditional writing assessments have often relied on artificial prompts disconnected from genuine communicative purposes—asking students to "write an essay about" topics they have no investment in, for audiences who will only evaluate rather than engage with ideas, in formats serving no purpose beyond demonstrating competence (White, 1994). Such decontextualized tasks may measure certain writing-related abilities but demonstrate questionable construct validity for predicting or representing performance in authentic situations where writing serves real communicative, epistemic, or professional functions (Lewkowicz, 2000). Students who struggle with artificial essay prompts might write effectively when addressing genuine rhetorical problems, they care about, while strong test performers may lack capabilities for navigating complex, real-world writing demands.

Authentic task design begins with careful analysis of actual writing situations students encounter or will face in academic, professional, or civic contexts. This analysis examines rhetorical features including purposes (to inform, persuade, analyze, synthesize, document, reflect), audiences (specialists, general readers, decision-makers, peers), genres (reports, proposals, analyses, arguments, narratives), and constraints (time availability, resource access, collaboration opportunities) characterizing consequential writing (Bachman & Palmer, 1996). Assessment tasks demonstrating authenticity incorporate these elements meaningfully rather than merely simulating surface features while maintaining artificial core purposes.

Several design principles enhance task authenticity. First, prompts should present genuine rhetorical problems requiring purposeful communication rather than display of writing skill in abstract. Instead of "write an essay about environmental issues," an authentic task might position students as interns preparing policy recommendations for local government based on environmental data, with specific audience needs and decision-making context (Wiggins, 1998). Such framing provides meaningful purpose, authentic audience, and realistic constraints that shape writing choices as they would beyond assessment situations.

Second, authentic tasks often integrate writing within broader activities rather than isolating composition as standalone skill demonstration. Research writing assessments might include locating and evaluating sources, synthesizing information across texts, and acknowledging source contributions—mirroring authentic academic writing processes (Howard, Serviss, & Rodrigue, 2010). Professional writing tasks could embed composition within case scenarios requiring analysis, decision-making, and communication to stakeholders. Such integration enhances construct validity by assessing writing as it functions authentically—as tool for thinking, learning, and communicating rather than end in itself.

Third, authentic assessment accommodates realistic resources and processes. Real-world writers typically access dictionaries, style guides, previous documents, and colleague consultation while composing; they engage in planning, drafting, seeking feedback, and substantive revision over extended periods. Assessment tasks can enhance authenticity by permitting appropriate resource use, allowing adequate time for recursive processes, and potentially incorporating peer review or teacher consultation that mirrors authentic writing situations (Elbow & Belanoff, 1997). While such accommodation complicates standardization and raises authentication concerns, it strengthens ecological validity by reducing artificial constraints that distort performance.

Genre diversity constitutes another authenticity dimension. Academic writing encompasses varied forms including literature reviews, research reports, analytical essays, lab reports, and reflective pieces, each with distinct conventions and purposes (Hyland, 2003). Professional contexts involve correspondence, proposals, reports, documentation, and presentations. Valid assessment of writing competence should therefore include multiple genres rather than treating the five-paragraph essay or similar limited formats as universal proxies for writing ability (Hillocks, 2002). Portfolio assessments prove particularly well-suited for incorporating genre diversity across multiple authentic tasks.

However, authenticity involves inherent tensions with other validity concerns. Highly contextualized tasks may introduce construct-irrelevant variance if students possess differential background knowledge unrelated to writing ability (Messick, 1994). Tasks requiring specialized content understanding risk confounding writing

competence with domain knowledge, potentially disadvantaging students encountering topics for the first time. Additionally, extensive authenticity—such as long-term projects with multiple drafts and outside resources—complicates attribution, raising questions about whose writing is being assessed when substantial assistance occurs (Hamp-Lyons & Condon, 2000).

Practical constraints also limit authenticity. Many authentic writing situations involve lengthy timelines, specialized tools, workplace access, or collaborative processes difficult to replicate in assessment contexts. Achieving complete authenticity proves neither feasible nor necessarily desirable if doing so compromises other validity aspects or practical implementation. The goal should be meaningful approximation of authentic tasks' essential cognitive, rhetorical, and linguistic demands rather than perfect simulation of every contextual detail (Cumming, 2002).

Ultimately, authentic task design represents not an absolute standard but a directional principle—assessment tasks should approximate genuine writing situations as closely as feasible while balancing competing validity considerations and practical constraints. Strategic selection of tasks that capture authentic purposes, audiences, and rhetorical challenges while remaining assessable within available resources provides optimal approach. Research examining relationships between authentic assessment performances and subsequent real-world writing success can provide validity evidence supporting or refuting inferences from more or less authentic tasks (Kane, 2013). By systematically pursuing authenticity as a design priority while acknowledging its limitations, writing assessment can better serve its fundamental purpose: supporting valid inferences about students' capabilities for the writing that genuinely matters in their academic, professional, and civic lives.

2.5 Writing Assessment in Libya

The study of writing assessments in higher education, particularly in the context of English as a Foreign Language (EFL), has been the focus of various empirical studies. These studies explore different methods of assessment, the challenges students and educators face, and the role of self-assessment in improving writing proficiency. The findings offer valuable insights into how assessments are conducted

and perceived by both students and instructors, shedding light on issues related to validity and fairness in writing assessments.

Waragh's (2016) study examines the methods of writing assessment used by EFL tutors in Libya, the factors influencing their practices, and students' perceptions of these assessments. Through surveys and interviews with tutors and students, the study finds that while both formative and summative assessments are employed, self-assessment and peer assessment are less common. A significant issue is the absence of clear assessment criteria, leaving students uncertain about what is expected from them. The study calls for more transparent assessment criteria and increased student participation in the feedback and grading process to enhance the writing assessment practices in Libya and other similar EFL contexts. A more recent study by Ramadan & Dekheel (2020) focuses on Libyan students' perceptions of traditional exams as an assessment method at Sirte University. This study reveals that students consider the current examination system unfair, unreliable, and outdated. They argue that it promotes rote memorization rather than evaluating their true abilities, which calls for reforms to improve the fairness and effectiveness of assessments in the Libyan higher education system. This highlights the need for more authentic assessment methods that better align with students' actual learning. That is, the need for more self-assessment practices.

Similarly, a study done by Eswaey & Ihmoumah (2024) about the role of self-assessment in EFL Writing development shows that self-assessment can significantly improve students' self- awareness, engagement, and overall writing performance. By regularly evaluating their own work, students develop a better understanding of their strengths and weaknesses, which leads to continuous improvement. Despite its potential benefits, the study points out that integrating self-assessment into EFL instruction presents challenges that require institutional and educator support. Dweni's (2023) qualitative study investigates the challenges faced by Libyan EFL undergraduate students in research writing. The study analyzed ten research papers and identified difficulties in areas such as building arguments, conducting critical analysis, referencing, and in-text citation. While issues related to coherence, grammar, and vocabulary were less problematic, the study suggests that a more structured research methods course-spanning two semesters and integrating both

theory and practice_ could significantly help students improve their research writing skills.

These studies examined the assessment leaving the area of the validity of this assessment. That is my research hope to shed more light on similarly to one study to the best of my knowledge by Dobrić's (2018) work which provides a theoretical and historical examination of the evolution of writing assessment validity, particularly in the field of language evaluation. Beginning in the 1920s and evolving through the late 1980s and early 1990s, Dobrić traces the development of test validation processes and their influence on educational theory and practice. The study explores how shifts in theoretical frameworks and real-world needs have impacted the validation of writing assessments. It also highlights the ongoing tension between the practices of higher education writing instructors and externally administered standardized exams. The work concludes by examining contemporary evaluation procedures and their continued evolution, particularly within EFL environments, where discussions on test validity continue to shape language education policies. Still this area needs more investigation that's where my study aims to add more insights to this topic. That is, the validity of the writing assessments**.**

2.6 Distinguishing the Current Study from Previous Research

While the reviewed literature provides valuable insights into writing assessment practices in EFL contexts, the current study distinguishes itself through its specific focus and methodological approach. Waragh's (2016) investigation examined tutors' assessment methods and student perceptions broadly, while Ramadan and Dekheel (2020) focused on students' views of traditional examinations as unfair and outdated. Eswaey and Ihmoumah (2024) explored self-assessment as a developmental tool, and Dweni (2023) identified specific challenges in research writing among Libyan undergraduates. Although Dobrić's (2018) theoretical work traced the historical evolution of writing assessment validity in language evaluation, it remained primarily conceptual rather than empirical.

In contrast, the present study directly investigates the validity of current EFL writing assessment practices at Sabratha University, specifically examining whether existing assessment instruments accurately measure students' writing competence across

multiple validity dimensions (content, construct, criterion, face, consequential, and ecological validity). Rather than focusing solely on perceptions, methods, or specific writing sub-skills, this research employs a systematic validity framework to evaluate the alignment between assessment practices and theoretical principles of effective writing evaluation. By empirically analyzing actual assessment tools, scoring procedures, and their correspondence with established validity criteria, this study addresses a critical gap in the Libyan EFL context where, despite growing awareness of assessment challenges, rigorous validity investigations remain scarce. Thus, while previous studies have illuminated various aspects of writing assessment and student experiences, the current research uniquely contributes by providing empirical evidence regarding the technical quality and appropriateness of writing assessments used in Libyan higher education, offering concrete recommendations for improving assessment validity and, consequently, educational outcomes.

2.7 Summary of the Chapter

This chapter provided a comprehensive review of EFL writing assessment, examining its theoretical foundations, practices, and validity concerns. Writing was explored as a complex skill integrating linguistic, cognitive, and sociocultural dimensions, encompassing micro-skills (grammar, spelling, structure) and macro-skills (coherence, organization, rhetorical awareness). The chapter outlined major assessment types—formative, summative, diagnostic, and self-assessment—emphasizing their distinct purposes in supporting learning and informing instruction. Key principles examined included reliability, validity, authenticity, practicality, and washback. Various methods were discussed, including holistic and analytic scoring, portfolio assessment, and performance-based assessment, with rubrics identified as essential tools for transparency and fairness. Validity emerged as the cornerstone of credible assessment, evolving from traditional content and construct validity to encompass consequential and ecological dimensions. Empirical studies from the Libyan context revealed ongoing challenges including assessment fairness, clarity of criteria, and limited self-assessment use. Traditional examinations often prioritize memorization over authentic writing competence, highlighting the need for reform toward reflective, student-centered models. The chapter concluded that effective EFL writing assessment requires balancing technical rigor with pedagogical purpose,

supporting the current study's investigation of assessment validity and its implications for improving teaching and learning in Libyan higher education.

Chapter Three: Methodology

3.0 Introduction

This chapter provides a detailed account of the methodology chapter. First it starts with the research design that the study followed. Then it highlights the participants and the tools used in this study. The chapter also gives a clear background of the methods used and the techniques of analyzing these methods. A triangulation of these methods was described and the pilot study to test the reliability of the methods. Finally, a summary of the chapter was set to sum up the main ideas covered.

3.1 Research Design

This study employs a mixed-methods approach, combining both quantitative and qualitative data collection tools to examine the validity of writing assessments in Libyan higher education. Specifically, in Sabratha University, Faculty of Education, Zoltun. The research utilizes questionnaires and document analysis of writing tests. According to Bryman (2012), Creswell (2015), and Creswell & Plano Clark (2011), mixed-methods research (MMR) is a comprehensive methodology that integrates different research approaches to address research questions in a methodologically rigorous and ethically sound manner. This approach allows for the collection, analysis, interpretation, and reporting of both quantitative and qualitative data, thereby providing a more comprehensive understanding of the assessment practices and their validity in evaluating students' writing proficiency.

3.1.1 Quantitative Approach:

The quantitative component of the mixed-methods approach involves collecting and analyzing numerical data to examine the breadth of the research question. In this study, questionnaires serve as the primary quantitative instrument, enabling the systematic collection of measurable data regarding writing assessment practices in Libyan higher education. This approach allows researchers to identify patterns, measure frequencies, and establish statistical relationships across a larger sample, providing generalizable insights into the validity perceptions and assessment practices employed by educators and experienced by students.

3.1.2 Qualitative Approach:

The qualitative component focuses on gathering narrative and descriptive data to provide depth and contextual understanding of the research problem. Through document analysis of writing tests, this approach examines how assessments are structured, what they measure, and how they are interpreted in practice. Qualitative methods enable researchers to explore the nuanced aspects of writing assessment validity that cannot be captured through numbers alone, offering rich insights into whether assessments truly measure students' writing abilities as intended and revealing the underlying complexities of assessment design and implementation in the Libyan higher education context.

3.1.3 Integrated Mixed-Methods Approach:

The mixed-methods approach (MMA) integrates both quantitative and qualitative research methods within a single study to provide a comprehensive perspective on the research problem. According to Creswell and Plano Clark (2011), this methodology combines numerical and narrative data to capitalize on the strengths of both approaches, enabling researchers to triangulate findings and offer richer, more robust insights than relying on a single method alone (Bryman, 2012; Johnson & Onwuegbuzie, 2004).

3.2 Sequential Design

In particular, this study uses a sequential mixed-methods design, in which data is gathered and analyzed in separate stages, with the results of one phase influencing the design of the next. In this study, the quantitative phase (the administration of questionnaires) is undertaken first, followed by the qualitative phase (document analysis of writing tests). According to Creswell (2015), a sequential design contains two main stages: a first stage where quantitative data is collected and analyzed followed by a second phase where qualitative data is obtained to further examine the findings. In order to collect quantifiable data on students' and instructors' opinion regarding the validity of writing assessments, the first phase of this study involves sending questionnaire to both groups. In the second phase, the qualitative analysis of writing test documents to assess their content, construct, and criterion validity.

This sequential approach allows the initial quantitative findings to guide the qualitative phase by identifying patterns and areas that need deeper exploration (Tashakkori & Teddlie, 2003). By following this design, the study not only assesses the general validity of writing assessments but also provides a nuanced understanding of their effectiveness and areas for improvement.

3.3 Sampling and Participants

The study's participants were chosen using a convenience sampling technique. This strategy, which is frequently used in educational research when random selection may not be practical, was selected since it was accessible and participants were willing to participate in the study. The study involved thirty students of Faculty of Education at Zulton-Sabratha University. The number was only thirty because that was the total number of the students in the faculty. The number of the teacher was six teachers who teach in the English Department at this Faculty.

3.4 Data Collection Tools

Two tools were used to collect data. Questionnaire which was used with both teachers and students and document analysis of previous writing tests. Both tools were used in order to gather a more comprehensive understanding of the topic.

3.4.1 Questionnaire

3.4.1.1 Teachers' Questionnaire:

The study employed a structured questionnaire for teachers to gather quantitative data regarding the validity of writing assessments used in the Faculty of Education at Zultun–Sabratha University. The questionnaire content was closely aligned with the official Aims and Objectives of the Writing Courses (Writing 1–Writing 5) taught in the English Department. Since these courses form the foundation upon which writing assessment practices are built, integrating their stated learning goals ensured that the questionnaire accurately captured whether current assessments measure the intended writing skills and competencies. The teachers' questionnaire examined their perceptions of how well writing assessments reflect the syllabus requirements, including: students' ability to write meaningful and grammatically correct sentences

(Writing 1), mastery of paragraph structure, unity, and coherence (Writing 2), competence in producing short and extended essays across rhetorical modes such as descriptive, narrative, argumentative, comparison/contrast, and cause–effect (Writing 3), proficiency in academic writing skills including paraphrasing, summarizing, referencing, interpreting data, and avoiding plagiarism (Writing 4), and familiarity with creative writing processes and techniques (Writing 5). Teachers were asked to evaluate whether existing writing assessments appropriately measure these targeted skills, whether assessment tasks match course objectives, and whether the assessment criteria ensure validity and reliability.

3.4.1.2 Students' Questionnaire:

The students' questionnaire was designed to reflect the writing skills they are expected to acquire throughout the five writing courses. Items probed students' self-evaluation of sentence writing, paragraph development, essay structure, use of rhetorical modes, revision strategies, academic writing abilities, and confidence in writing—core learning outcomes emphasized across the writing curriculum. This ensured that the questionnaire captured students' perceptions of how well assessments represent what they were taught and whether the tests allow them to demonstrate these skills effectively. By grounding the instrument in the documented aims and objectives of the writing program, the study ensured that the questionnaire was valid, context-appropriate, and directly aligned with the competencies that writing assessments are expected to measure within Libyan EFL higher education.

The selection of questionnaires as research tools was based on their suitability for examining perceptions of assessment practices across different stakeholders. Convenience sampling was used due to ease of access and participants' willingness to take part in the study (Etikan, Musa, & Alkassim, 2016; Creswell, 2012).

3.4.2 Document Analysis of Previous Writing Tests

Document analysis is a methodical process of examining or assessing documents, both printed and electronic, that requires the systematic examination and interpretation of data to extract meaning, understand phenomena, and generate empirical knowledge (Corbin & Strauss, 2008; see also Rapley, 2007). It is often used alongside other research methods as a means of triangulation to strengthen

findings. Document analysis is an efficient and cost- effective qualitative research method, offering advantages such as quick data collection, the availability of documents in the public domain, and minimal obtrusiveness and reactivity (Merriam, 1988). Many documents are accessible without needing the author's permission, allowing researchers to gather data without influencing the original context or content. However, limitations such as insufficient detail, low retrievability, and biased selectivity must be acknowledged; these concerns can be mitigated by carefully selecting documents that are produced for specific purposes and aligned with organizational policies and procedures, making the disadvantages relatively minor compared to the method's practical benefits (Yin, 1994). Document analysis is inherently an iterative process involving skimming, reading, and interpretation, and it incorporates elements of both content analysis and thematic analysis by organizing information into categories based on the study's central research questions. While some researchers may objectify content analysis as merely a preliminary review to separate relevant from irrelevant information, thematic analysis emphasizes identifying patterns within the data, with emerging themes forming the foundation for deeper analysis. Throughout this process, the researcher must engage in careful re-reading, coding, and category construction to uncover themes pertinent to the phenomenon under investigation, using predefined codes when appropriate, especially if document analysis serves as a supplementary method. Ultimately, the researcher must maintain objectivity and sensitivity when selecting and interpreting documents to ensure the credibility and trustworthiness of the findings.

3.5 Triangulation

Triangulation is a strategy used to enhance the credibility and validity of research findings by combining multiple methods or data sources. According to Patton (1990), triangulation allows researchers to avoid the criticism that a study's conclusions are merely the result of biases stemming from a single researcher, source, or method. By using different forms of data collection, the researcher can strengthen the trustworthiness of the results. Eisner (1991) also emphasizes that triangulation provides "a confluence of evidence that nurtures trust," helping to build a more comprehensive and reliable understanding of the phenomenon under investigation.

In this study, triangulation was achieved through the use of both questionnaires and document analysis of previous writing tests. By collecting data through structured questionnaires from students and teachers, and by systematically analyzing samples of students' past writing assessments, the researcher was able to validate findings across different data sets. This approach helped to minimize potential biases inherent in any single method and provided a richer, more nuanced picture of the validity of writing assessments in the studied context.

3.6 Piloting the Questionnaire

Before implementing any research method, it is essential to assess its feasibility, clarity, and usability. As Wallen and Fraenkel (2001) state, these factors must be carefully considered to ensure the effectiveness of the chosen instruments. Therefore, a pilot study was conducted prior to the main data collection to test the research design and identify any potential issues. The questionnaire was piloted with a small group consisting of two teachers and five students. The feedback from the participants indicated that the questions were clear and understandable, and the time required to complete the questionnaire was approximately twenty minutes, confirming the practicality of the instrument for the larger study. The results of the pilot study demonstrated that the questionnaire was both feasible and user- friendly. Participants reported no difficulties in understanding the wording or structure of the questions, suggesting that the content was appropriately designed for the target audience. Additionally, the time taken to complete the questionnaire was reasonable, which minimized the likelihood of respondent fatigue in the main study. Based on the pilot study findings, no major modifications were necessary, and the questionnaire was deemed ready for full-scale administration.

3.7 Ethical Issues

This study carefully followed ethical guidelines to ensure the protection, dignity, and rights of all participants, in line with the principles set out by the British Educational Research Association (BERA, 2018). Prior to administering the questionnaire, participants were fully informed about the aims, procedures, and voluntary nature of the study. Informed consent was obtained from all participants — both teachers and students — with clear explanations that they could withdraw at any time without

giving a reason and without any negative consequences.

The questionnaire was administered in a manner that respected participants' autonomy, privacy, and confidentiality. No personal identifying information was collected, ensuring participants' anonymity. Participants were assured that the data collected would be used solely for academic purposes and that their individual responses would not be disclosed or traced back to them.

In terms of reporting results, data were analyzed and presented in a way that maintained confidentiality and avoided any risk of identifying individual participants or institutions. The findings were reported honestly, without fabrication, misrepresentation, or selective reporting of results.

Permission to conduct the study was obtained from the relevant academic authorities at Faculty of Education at Zulton, Sabratha University, ensuring institutional support and ethical approval. Throughout the research process, care was taken to treat all participants with respect and to uphold the standards of fairness, transparency, and accountability expected in British educational research.

3.8 Summary of the chapter

The methodology chapter of the study was structured around key components, starting with the research design, which employed a mixed-methods approach combining both quantitative and qualitative data collection methods. The study followed a sequential design, with data being collected in distinct stages, starting with questionnaires for both students and teachers, followed by document analysis of writing tests. Participants were selected using convenience sampling, involving 30 students and 6 teachers from the Faculty of Education at Sabratha University. To ensure the reliability and validity of the methods, triangulation was employed, combining different data sources and research tools. A pilot study was conducted to test the questionnaires' clarity and feasibility, which demonstrated their suitability for the main study. Ethical considerations were carefully addressed, with participants being fully informed, their consent obtained, and their anonymity and confidentiality respected throughout the study. The chapter concluded with a summary of the research methods used, emphasizing the systematic approach to gathering and analyzing data.

Chapter Four: Data Analysis and Findings

4.0 Introduction

As mentioned in chapter three, this study adopted a mixed-methods approach. In other words, quantitative and qualitative principles were used, considering that combining both types of methods enables a better understanding of the problems and situations under analysis (Cresswell & Clark, 2011). A mixed-methods methodology allows for the combination of different types of information in a single study and may provide solutions to issues that could arise when a single method is used (Cresswell et al., 2003, cited in Glowka, 2011; Creswell, 2013).

Data for this study was obtained from the ten test paper samples scored by the 6 EFL teachers as well as a close- ended questionnaire delivered to both participating teachers and students**.**

4.1 Data Analysis

The analysis of data was carried out in two main phases. In phase one, the answers to the background questionnaire were analyzed in an attempt to understand the grading participants' perceptions towards writing assessment and assessment tools. Answers were grouped into contrasting categories in order to obtain a general perspective on the information. In the second phase, data obtained from the scores for each sample were entered into a statistics software program (SPSS) with the purpose of obtaining descriptive statistics such as the mean (M) and standard deviation (SD), and a t-test was performed to compare the M and SD obtained and identify significant differences. The sum of the five analytical scores given to each paper was considered for this statistical analysis. The calculations obtained were then compared between one paper and another and between one teacher and the other to identify important differences and similarities in the data. With the purpose of ensuring the validity of the data obtained, information was discussed among the authors of this study, and then independently with an external expert researcher.

The questionnaire comprised closed-ended multiple-choice items that facilitated systematic data collection and analysis while providing participants with structured opportunities to express their perspectives. To ensure the instrument's clarity and

effectiveness, a pilot study was conducted with a group of English language teachers who were not included in the main research sample. This preliminary testing aimed to gather feedback regarding the questionnaire's comprehensibility, relevance, and alignment with its intended purpose, allowing the researcher to refine the instrument before its administration to the actual study participants.

4.1.1 Analysis of Background Questionnaire

The participants in this study answered a background questionnaire that elicited their general background, teaching experience, and perceptions towards writing assessment and analytical scoring rubrics. The results are presented in Table (1)

Table (2): Results of background questionnaire

Variable		N	%
Gender	Male	7. 5	8. 83%
	Female	10. 1	11. 16%
Experience	1-4 years	14. -	15. -
	10-14 years	17. 3	18. 50%
	20-24 years	20. 3	21. 50%
Level of education	Master	24. 3	25. 50%
	Doctorate	27. 3	28. 50%
		31.	32.

Table (2) revealed that the majority of the participants (83%) are males, while females constitute only 16% of the participants. As for the experience, it demonstrated that none of the participants has work experience in the range of 1-4 years. Half of the participants (50%) have work experience in the range of 10-14 years, and another 3 participants (50%) have work experience in the range of 20-24 years. The level of education is also shown in the same table and it reveals that three participants (50%) have a master's degree, and three participants (50%) have a doctorate degree.

4.1.2 Analysis of Teachers' Questionnaire

In this section, the ten questions in the teachers' questionnaire are analyzed one by one. The results are tabulated and followed by a brief explanation for each one.

1. How frequently do you administer writing assessments in your EFL classroom?

This question investigates how often teachers do assessment in their classrooms.

Table (3): Frequency of administering writing assessments

Item 1	Rarely	Occasionally	Monthly	Weekly	M	SD
How frequently do you administer writing assessments in your EFL classroom?	**-**	**1**	**2**	**3**	**3.33**	**8.17**
		17%	**33%**	**50%**		

Table (3) revealed that half of the participants (50%) administer writing assessments in their EFL classroom weekly, while 33% administer writing assessments monthly. One teacher (17%) administers writing assessments occasionally. However, none of the participants administer writing assessments rarely. In general, the frequency of administering writing assessments seems to be at regular level [M=3.33, SD=.817].

2. How do you define the purpose of EFL writing assessments in your classroom?

The purpose of this question is to understand the reasons why teachers do assessment.ie. for assessing language proficiency, achievement or evaluating students' understanding.

Table (4): The purpose of EFL writing assessment in your classroom

Item 2	Assessing language proficiency	Promoting language development	Evaluating students' understanding of writing conventions	M	SD
How do you define the purpose of EFL writing assessment in your classroom?	**3**	**1**	**2**	**1.8**	**.98**
	50%	**17%**	**33%**		

Table (4) shows that the majority of the teachers (50%) stated that assessing the language proficiency is the purpose of writing assessment. 33% of the participants consider evaluating students' understanding of writing conventions is the purpose

of writing assessment, while only 17% believe that the purpose of writing assessment is promoting language development.

3. **What types of writing prompts do you typically use in EFL writing assessments?**

The aim of this question is to understand the type of the assessment used such as whether it is Opinion, descriptive, narrative or expository.

Table (5): Types of writing prompts

Item 3	Opinion/ argumentative	Descriptive	Narrative	Expository	M	SD
What types of writing prompts do you typically use in EFL writing assessments?	**3**	**1**	**2**	**-**	**1.6**	**0.82**
	50%	**17%**	**33%**			

As shown in Table (5) most of the teachers (50%) use opinion or argumentative prompts in their writing assessment, while 33% use narrative prompts. Only one teacher (17%) uses descriptive prompts and none of the participants use expository prompts in their writing assessments.

4. **How do you ensure that the writing prompts used in assessments are relevant and meaningful to your EFL students?**

The meaning of this one is to reveal whether teacher use aligning prompts with students' interests and experiences, incorporating real-world contexts, or connecting prompts to the curriculum to make their assessment meaningful. This is shown in the following table.

Table (6): Relevance of writing prompts

Item 4	Aligning prompts with students interests	Incorporating real world context	Connecting prompts with curriculum	M	SD

How do you ensure that the writing prompts used in assessments are relevant and meaningful to your EFL students?	**1**	**2**	**3**	**2.33**	**0.82**
	17%	**33%**	**50%**		

From Table (6) it can be observed that 50% of the participants connect prompts with curriculum in order to ensure that the writing prompts used in assessments are relevant and meaningful to EFL students, and 33% incorporate real world context to ensure the relevance of writing prompts. Only one teacher (17%) aligns prompts with students' interests to ensure that the writing prompts used in assessments are relevant and meaningful to EFL students.

5. How do you assess the validity of EFL writing assessments in your classroom?

By this question, the researcher wanted to understand the way teachers usually use to assess their students. For example, whether they compare students' performance with established writing standards, use rubrics to evaluate specific writing criteria (e.g., grammar, vocabulary, organization), or analyze the correlation between writing scores and other language proficiency measures (e.g., speaking, listening).

The answers for this question are analyzed in the following table.

Table (7): Assessing the validity of EFL writing assessments

Item 5	Comparing students' performance with established writing standards	Using rubrics to evaluate specific writing criteria	Analyzing the correlation between writing scores and other language proficiency measures	M	SD
How do you Assess the validity of EFL writing assessments in your classroom?	**3**	**3**	**-**	**2.5**	**0.22**
	50%	**50%**			

It is apparent, from Table (7) that half of the participants (50%) assess the validity of EFL writing assessments in the classroom through comparing students' performance with established writing standards, while the other half (50%) assess the validity of EFL writing using rubrics to evaluate specific writing criteria. On the other, none of the participants assess the validity of EFL writing by analyzing the correlation between writing scores and other language proficiency measures such as listening and speaking.

6. How do you address the issue of reliability in EFL writing assessments?

Using multiple raters to evaluate students' writing or offering opportunities for students to revise and improve their writing are examples of addressing the issue of reliability in EFL writing assessments.

The analysis of this issue of reliability is shown in Table below.

Table (8): The issue of reliability

Item 6	Using multiple raters to evaluate students' writing	Providing clear and consistent scoring criteria	Offering opportunities for students to revise and improve their writing	M	SD
How do you address the issue of reliability in EFL writing assessments?	**2**	**3**	**1**	**1.8**	**0.75**
	33%	**50%**	**17%**		

Table (8) revealed that half of the participants (50%) provide clear and consistent scoring criteria to assess the reliability of EFL students' writing, while 33% of the participants use multiple raters to evaluate students' writing. A small percentage of the participants (17%) offer opportunities for students to revise and improve their writing. On average, the participants seem to use different methods to assess the reliability of their assessment of students' writing [M=1.8, SD=0.75].

7. **In your opinion, what are the biggest challenges in assessing EFL writing validity?**

There a r e many challenges in assessing writing as mentioned in the review chapter, but the researcher wanted to investigate the biggest challenges. Here, some suggestions have been provided for the teachers such as differentiating between language errors and developmental stages, capturing the complexity of language use in a single assessment, and/or balancing the need for accuracy and fluency. If there are any other challenges, the participants are allowed to add any.

Table (9) displays the results of these challenges in assessing EFL writing validity.

Table (9): Challenges in assessing EFL writing validity

Item 7	Differentiating between language errors and developmental stages	Capturing the complexity of language in a single assessment	Balancing the need for accuracy and fluency	M	SD
In your opinion, what are the biggest challenges in Assessing EFL writing validity?	**2**	**1**	**3**	**2.2**	**0.98**
	33%	**17%**	**50%**		

It can be observed that the majority of the teachers (50%) consider balancing the need for accuracy and fluency to be the biggest challenges in assessing EFL writing validity, while 33% of the teachers see differentiating between language errors and developmental stages to be the biggest challenges. However, only 17% consider capturing the complexity of language in a single assessment as the biggest challenges in assessing EFL writing. On average, it seems that the teachers face some challenges in assessing EFL writing [M=2.2, SD= 0.98].

8. **How do you provide feedback to students based on their writing assessments?**

Simply, this question is on the way teachers provide the feedback. For instance, teachers might write comments on specific strengths and areas for improvement,

one-on-one conferences to discuss their students' writing, or they might just do peer feedback. If there are any other ways, enough space is provided for any comments from the participants.

Table (10) presents the results of providing feedback to students.

Table (10): Providing feedback

Item 8	Written comments On specific strength.	One on one conference to Discuss their writing.	Peer feedback and revision activities.	M	SD
How do you provide feedback to students based on Their writing assessments?	**1**	**2**	**3**	**2.33**	**0.82**
	17%	**33%**	**50%**		

Table (10) shows that most of the teachers (50%) provide feedback to the students in their writing through peer feedback and revision activities. On the other hand, 33% of the participants provide feedback to the students through conducting conferences with the students to discuss their writing, while only 17% of the participants provide feedback to the students through written comments on specific strength. Hence, it can be concluded that the teacher keeps providing feedback on the students' writing [M=2.33, SD=0.82].

9. **How do you use the results of EFL writing assessments to inform your instruction?**

The instruction can be given to the students in many ways. Identifying individual student needs and designing targeted instruction, adjusting the pace and content of the curriculum or providing additional support or enrichment opportunities are among the common ways. These ways are provided for the participants to tick the right one and, as usual, a space is provided for further comments.

Table (11) presents the results of using the results of EFL writing assessment to inform instruction.

Table (11): Use of the results to inform instruction

Item 9	Identifying individual student's needs.	Adjusting the content of the curriculum	Providing additional support.	M	SD
Use of the results of EFL writing Assessment to inform instruction	**3**	**1**	**2**	**1.83**	**0.98**
	50%	**17%**	**33%**		

It can be noted that the majority of the teachers (50%) of the participants use the result of the writing assessment to inform instruction by identifying individual students' needs and designing targeted instructions, while 33% inform instructions by providing additional support. Only one teacher (17%) informs instructions by adjusting the pace and content of the curriculum. On average, it can be concluded that the teachers use the results of EFL writing assessment to inform instruction at a moderate level [M=1.83, SD=0.98].

10. What changes or improvements would you suggest to enhance the validity of EFL writing assessments?

This question is about teachers' suggestions and recommendations to improve the validity of EFL writing assessments.

Table (12): Suggested Improvements to enhance the validity

Item 10	Including a wide range of writing genres	Incorporating authentic writing tasks	Providing more opportunities for students to engage in writing	M	SD
What changes or improvements would you suggest to enhance the Validity of EFL Writing	**1**	**1**	**4**	**2.5**	**0.83**
	17%	**17%**	**67%**		

assessments?					

Table (12) revealed that more than half of the participants (67%) suggest providing more opportunities for students to engage in writing in order to improve the validity of EFL writing assessments. However, only one teacher (17%) suggests including a wide range of writing genres, and another one teacher (17%) suggest incorporating authentic writing tasks in order to improve the validity of EFL writing assessments. It can be stated that the teachers provide some suggestions to improve the writing assessment at a high level [M=2.5, SD=0.83].

4.1.2.1 Summary of Teachers' Questionnaire Findings

Teachers regularly administer writing assessments, with half conducting them weekly and primarily focusing on evaluating language proficiency through opinion/argumentative prompts connected to curriculum standards. They assess validity by comparing student performance with established standards or using rubrics, while ensuring reliability through clear scoring criteria or multiple raters. The main challenge identified is balancing accuracy and fluency in assessment, and teachers predominantly provide feedback through peer feedback and revision activities. Most teachers use assessment results to identify individual student needs for targeted instruction, and they suggest that providing more opportunities for students to engage in writing would significantly enhance the validity of EFL writing assessments.

4.1.3 Analysis of Students' Questionnaire

This analysis aims to evaluate students' perception about their English writing skills through a structured questionnaire. The data collected offers insights into the varying levels of self- assessment among students.

1. How confident do you feel about your writing skills in English?

This statistical report presents the results of a survey conducted to assess the confidence levels of individuals regarding their writing skills in English. By this question the researcher seeks to know the mental state of students while writing.

Table (13) presents the descriptive analysis of the level of confidence of the students' writing skills.

Table (13): the students' confidence on writing

Item 1	Very confident	Somewhat confident	Not very confident	M	SD
How confident do you feel about your writing skills in English?	**5**	**10**	**5**	**2**	**0.73**
	25%	**50%**	**25%**		

The participants' responses indicate varying levels of confidence in their writing skills in English. Out of the 20 participants, 5 (25%) indicate that they are "Very Confident" in their writing skills, 10 (50%) responded as "Somewhat Confident", and 5 (25%) expressed being "Not Very Confident" in their writing skills. In general, it can be stated that the participants have a moderate level of confidence in their writing [M=2, SD=0.73].

2. How often do you practice writing in English outside of the classroom?

This question asks about the amount of practice of English writing outside the classroom's session. This section presents the findings of a survey conducted to determine the frequency

at which individuals practice writing in English outside of the classroom. The results are displayed in Table (14)**.**

Table (14): Frequency of practicing writing outside the classroom

Item 2	Daily	Several times at week	Once a week	Rarely	Never	M	SD
Frequency of practicing writing outside the classroom	**3**	**4**	**5**	**6**	**2**	**3**	**1.26**
	15%	**20%**	**25%**	**30%**	**10%**		

Table (14) revealed that 30% of the participants rarely practice writing outside the classroom, 25% practice writing once a week, and 20% practice writing several times at week. On the other hand, only 15% of the students practice writing daily, while 10% do not practice writing. This suggests that the participants, on average, engage in writing practice in English outside of the classroom at a moderate frequency [M=3, SD=1.26].

3. Which types of writing tasks do you find most challenging?

This question examines the types of writing tasks that might be difficult on students' ability. This item presents the results of a survey conducted to identify the types of writing tasks that students find most challenging. The results are shown in Table (15).

Table (15): Most challenging types of writing task

Item 3	essay writing	letter writing	creative writing	report writing	M	SD
Most challenging types of writing tasks	14	1	2	3	**1.7**	**1.17**
	70%	5%	10%	20%		

It can be noted, from Table (15) that most of the participants (70%) reported that essay writing is the most challenging type of writing task. One participant (5%) consider letter writing is the most challenging task. Two participants (10%) chose creative writing, and three participants (15%) indicated report writing as their most challenging task. This suggests that, on average, the participants found the selected types of writing tasks moderately challenging [M=1.7, SD=1.17].

4. On a scale of 1-5, how well do you understand the structure and organization of a well- written paragraph or essay?

The question aims to find out the amount of difficulty structure and organization are on students' ability. This section presents the findings of a survey conducted to assess individuals' understanding of the structure and organization of a well-written paragraph or essay.

Table (16): Understanding a well-written paragraph /essay

Item 4	Not at all	Scale three	Scale four	Very well	M	SD
How well do you understand the structure and organization of a well-written paragraph or essay?	**2**	**9**	**5**	**4**	**3.45**	**0.89**
	10%	**45%**	**25%**	**20%**		

Table (16) revealed that out of the 20 participants, 2 (10%) indicated that they have "Not at All" understanding of the structure and organization of a well-written paragraph or essay. Nine participants (45%) rated their understanding as a "three" on the scale, five participants (25%) rated it as a "four," and four participants (20%) indicated they understand it "Very Well. "The participants' responses indicate varying levels of understanding regarding the structure and organization of a well-written paragraph or essay. The overall results suggest a moderate to high level of understanding among the participants in relation to the structure and organization of a well-written paragraph or essay [M=3.45, SD=.89].

5. How comfortable are you with using grammar and vocabulary accurately in your writing?

This question dive more on students' knowledge of writing grammar and rules. This section presents the findings of a survey conducted to assess individuals' comfort level with using grammar and vocabulary accurately in their writing.

Table (17): using grammar and vocabulary accurately in writing

Item 5	very comfortable	somewhat comfortable	not very comfortable	M	SD
How comfortable are you with using grammar and vocabulary accurately in your writing?	**2**	**13**	**5**	**2.15**	**0.58**
	10%	**65%**	**25%**		

Table (17) showed that the majority of the participants (65%) feel somewhat comfortable with using grammar and vocabulary accurately in writing. 25% of the

participants are not very comfortable with using grammar and vocabulary accurately in writing, while a small percentage of participants (10%) are very comfortable. on average, the participants have a moderate to somewhat comfortable with grammar and vocabulary usage in writing [M=2.15, SD= 0.58).

6. How often do you seek feedback or assistance from your teacher or peers regarding your writing?

This question states the assistant effect of the teacher on students' ability of writing. The items reveal the statistics of the feedback of the teacher comments when students try to improve.

Table (18): Seeking feedback from the teacher or peers

Item 6	Always	Often	Sometimes	Rarely	Never	M	SD
How often do you seek feedback or assistance from your teacher or peers regarding your writing?	**1**	**2**	**14**	**2**	**1**	**3**	**0.79**
	5%	**10%**	**70%**	**10%**	**5%**		

Table (18) revealed that most of the students (70%) sometimes seek feedback or assistance from your teacher or peers regarding their writing. On the other hand, a small number of students (10%) indicated they are "often" or "rarely" seek feedback. Only one student (5%) always seeks feedback, while another student never seeks any feedback. This suggests that, on average, the participants seek feedback or assistance from their teacher or peers regarding their writing at a moderate level [M=3, SD= 0.79].

7. How well do you think you can express your ideas and thoughts clearly in writing?

This statistical report presents the findings of a survey conducted to assess individuals' perceptions of their ability to express their ideas and thoughts clearly in writing.

Table (19): Ability to express ideas and thoughts in writing

Item 7	Very well	Moderately well	Not very well	M	SD

How well do you think you can express your ideas and thoughts clearly in writing?	**11**	**7**	**2**	**1.5**	**0.69**
	55%	**35%**	**10%**		

Table (19) revealed that more than half of the participants (55%) think they can express their ideas in writing very well, while 35% of the participants think they can express their ideas moderately well. However, only 10% are not very well at expressing their ideas in writing. It can be suggested that the participants perceive themselves to be moderately proficient in expressing their ideas and thoughts clearly in writing [M=1.5, SD=0.69].

8. How frequently do you revise and edit your writing to improve its quality?

This statistical report presents the findings of a survey conducted to determine the frequency with which individuals revise and edit their writing to improve its quality.

Table (20): Revising and editing writing to improve its quality

Item 8	Always	Often	Sometimes	Rarely	M	SD
How frequently do you revise and edit your writing to improve its quality?	**8**	**2**	**9**	**1**	**2.2**	**1.03**
	40%	**10%**	**45%**	**5%**		

The results in Table (20) demonstrated that a significant percentage of the participants (45%) sometimes revise their writing to improve its quality, while 40% always revise and edit their writing. However, 10% of the students often revise and edit their writing and only 5% rarely revise and edit their writing. On average, the participants revise and edit their writing to improve its quality at a moderate level [M=2.2, SD=1.03].

9. How comfortable are you with using appropriate academic or formal language in your writing?

This statistical report presents the findings of a survey conducted to assess individuals' comfort level with using appropriate academic or formal language in their writing.

Table (21): using appropriate academic or formal language in writing

Item 9	Very comfortable	Somewhat comfortable	Not very comfortable	Not very comfortable at all	M	SD
How comfortable are You with using appropriate academic or formal language in your writing?	**2**	**9**	**8**	**1**	**2.4**	**0.75**
	10%	**45%**	**40%**	**5%**		

Table (21) shows that a great percentage of the students (45%) somewhat comfortable with using appropriate academic or formal language in their writing, while 40% are not very comfortable. On the other hand, only 10% feel very comfortable with the use of appropriate academic or formal, and 5% is not very comfortable at all. On average, it can be mentioned that the participants have a moderate level of comfort with using appropriate academic or formal language in their writing [M=2.4, SD=.75].

10. How would you rate your overall writing competence in English compared to your other language skills (listening, speaking, reading)?

This statistical report reveals the results f the survey conducted to rate the overall writing competence in English compared to other language students may possess.

Table (22): Overall writing competence

Item 10	stronger than any skill	the same as other skills.	weaker than other skill	M	SD
How would you rate your overall writing competence in English compared to your other language skills (listening, speaking, reading)?	**4**	**14**	**2**	**1.9**	**0.55**
	20%	**70%**	**10%**		

From Table (22) it can be observed that the majority of the students (70%) rate their

writing skill as the same as other skills, while 20% of the students rate their writing stronger than other skills. However, 10% of the participants rate their writing as weaker than other skills. On average, it can be concluded that the participants perceive their writing competence in English to be similar to their other language skills [M=1.9, SD=0.55].

11. In your opinion, what areas of writing do you need the most improvement in?

This statistical report presents the findings of the areas of writing that individuals believe they need the most improvement in.

Table (23): Areas of writing that need the most improvement

Item 8	Grammar	Vocabulary	Organization	Clarity of ideas	M	SD
what areas of writing do you need the most improvement in?	**6**	**11**	**1**	**2**	**1.95**	**0.88**
	30%	**55%**	**5%**	**10%**		

The results show that more than half of the participants (55%) identify grammar as the most area that need to be improved, while 30% consider grammar to be improved. Organization and clarity seem to perceive the least percentage for improvement as 10% of the students identify clarity of ideas to be improved and only 5% of the participants see organization as an area for improvement. Overall, it can be concluded that the participants perceive multiple areas of their writing that could be improved [M=1.95, SD=0.88].

12. How do you typically prepare or plan your writing before you start?

This statistical report represents the findings collected from students about their brainstorming ways before preparing to write.

Table (24): Planning writing

Item 12	Outlining or creating structure	Brainstorming ideas	Researching and gathering information	M	SD

How do you typically prepare or plan your writing before you start?	**2**	**8**	**10**	**2.4**	**0.68**
	10%	**40%**	**50%**		

The majority of participants (50%) stated that they typically prepare for their writing by researching and gathering information, while 40% reported that they engage in brainstorming ideas as part of their preparation process. A smaller proportion of participants (10%) engage in outlining or creating a structure. Overall, it is suggested that, on average, the participants combine multiple methods to prepare and plan their writing [M=2.4, SD=0.68].

4.1.3.1 Summary of Students' Questionnaire Findings

Students demonstrate moderate confidence in their English writing skills, with half feeling somewhat confident, though they practice inconsistently outside the classroom with 30% rarely practicing. Essay writing is overwhelmingly the most challenging task for 70% of students, while they show moderate to high understanding of paragraph structure and somewhat comfortable levels with grammar and vocabulary accuracy. Most students believe they can express ideas well in writing and revise their work regularly, yet only sometimes seek feedback from teachers or peers. Students rate their writing competence as equal to their other language skills, identify vocabulary as the primary area needing improvement followed by grammar, and typically prepare for writing by researching information or brainstorming ideas rather than creating structured outlines.

4.2 Document Analysis of Writing Test Papers

In this study, the analytical rubric was chosen as the primary method for evaluating writing papers due to its ability to provide a structured, objective, and reliable measure of key writing competencies such as grammar, coherence, and organization. The rubric's detailed breakdown of each aspect of writing allowed for a nuanced and consistent evaluation of student proficiency, aligning with the study's goal to assess the validity of current writing assessment practices. Weigle, 2002; Hamp-Lyons (1991). However, to complement the rubric’s structured approach, thematic analysis was employed to capture the more subjective, context-dependent

aspects of the writing, such as common themes in language use, feedback application, and student engagement with academic writing conventions Braun & Clarke (2006). This combination of quantitative and qualitative methods allowed for a deeper, more comprehensive understanding of the writing assessment process, revealing both measurable weaknesses in writing skills and the underlying challenges students face in applying feedback and improving their writhing Weigle,2002; Hamp-Lyons (1991).

4.2.1 Triangulation of both Thematic and Analytical Rubric for the Test Paper Analysis.

The combination of analytical rubric and thematic analysis was intentionally employed in this study to triangulate both quantitative and qualitative data, ensuring a comprehensive evaluation of the students' writing performance. The analytical rubric was first used to provide a structured, quantitative assessment of the students' writing, focusing on specific aspects such as grammar, vocabulary, organization, and coherence (Weigle, 2002). However, while the rubric offered a clear, objective measurement of writing quality, it did not capture the nuanced themes and patterns in the writing that could provide additional insights into students' skills and challenges. To address this limitation, thematic analysis was employed to explore recurring themes and patterns within the writing samples, especially those related to content, coherence, and the students' writing process. Braun & Clarke, (2006). This approach allowed for a richer, more interpretive understanding of the student's writing beyond what was quantifiable by the rubric alone.

4.2.1.1 Results and Analysis of EFL Writing Assessment Using the Analytical Rubric

Table (25): Statistical Overview and Interpretation of Test Paper 1

Assessment Criterion	Mean Score	Median	Standard Deviation
Language	3.00	3.00	-
Vocabulary	2.80	2.80	-
Organization	3.33	-	0.52
Mechanics	2.33	-	-

In-course Content	3.00	-	-

The first test paper reveals a moderately consistent performance across most writing dimensions. Language use demonstrates acceptable proficiency, with both mean and median values stabilizing at 3.00 on the five-point scale. Vocabulary deployment appears satisfactory, though slightly weaker at 2.80, suggesting students possess adequate lexical resources but may benefit from expanding their word choices.

Organization emerges as a relative strength, with a mean score of 3.33, though the standard deviation of 0.52 indicates some variability in how students structure their writing. The most concerning area appears to be mechanics, which achieved only 2.33, pointing to persistent challenges with grammar, punctuation, and spelling conventions. The content appropriateness score of 3.00 suggests that students generally understand the task requirements and can produce relevant responses.

Table (26): Statistical Overview and Interpretation of Test Paper 2

Assessment Criterion	Mean Score	Median	Standard Deviation
Language	2.83	3.00	-
Vocabulary	2.83	3.00	-
Organization	3.00	3.00	-
Mechanics	3.00	3.00	-
In-course Content	2.83	3.00	-

Test Paper 2 exhibits remarkable consistency across all measured dimensions, with scores clustering tightly between 2.83 and 3.00. This narrow range suggests a balanced, if somewhat modest, writing competence. The uniformity of median values at 3.00 across all categories indicates stable performance patterns among the student sample.

Organization and mechanics both reached 3.00, representing the strongest areas in this assessment. Meanwhile, language use, vocabulary, and content relevance scored marginally lower at 2.83, though these differences are practically negligible. The overall pattern suggests that students demonstrate adequate but unremarkable writing abilities, with neither pronounced strengths nor glaring weaknesses characterizing this particular test administration.

Table (27): Statistical Overview and Interpretation of Test Paper 3

Assessment Criterion	Mean Score	Median	Standard Deviation
Language	3.67	4.00	Low
Vocabulary	3.33	3.00	Low
Organization	4.00	4.00	Low
Mechanics	3.17	3.00	Low
In-course Content	3.17	3.00	Low

This test paper demonstrates notably stronger performance compared to the previous assessments. Organization stands out as particularly commendable, achieving a mean of 4.00 with consistent median values, suggesting that students successfully implemented coherent structural frameworks in their writing.

Language proficiency also improved substantially, reaching 3.67 with a median of 4.00, which indicates that most students demonstrated good command of linguistic features. Vocabulary, mechanics, and content all fall within the satisfactory to good range (3.17-3.33). The consistently low standard deviations across all dimensions represent a key finding, as they indicate homogeneous performance levels within the sample, minimizing the variation typically seen in language assessment contexts.

Table (28): Statistical Overview and Interpretation of Test Paper 4

Assessment Criterion	Mean Score	Median	Standard Deviation
Language	2.83	2.50	Relatively High
Vocabulary	3.33	3.00	Moderate
Organization	3.67	4.00	Moderate
Mechanics	3.00	3.00	Relatively High
In-course Content	3.17	3.00	Moderate

Test Paper 4 presents a more complex picture, with mixed results warranting careful interpretation. Organization performed well at 3.67, and vocabulary usage was satisfactory at 3.33. However, language use emerges as problematic, with a mean of 2.83 and a notably lower median of 2.50, suggesting that a substantial portion of students struggled with linguistic accuracy and appropriateness.

The relatively high standard deviations observed in language and mechanics dimensions raise concerns about consistency. This variability indicates that some students performed reasonably well while others faced considerable difficulties, pointing to potential disparities in language proficiency levels within the group. Such heterogeneity may necessitate differentiated instructional approaches to address the varying needs of learners.

Table (29): Statistical Overview and Interpretation of Test Paper 5

Assessment Criterion	Mean Score	Median	Standard Deviation
Language	3.50	3.50	Low
Vocabulary	3.33	3.00	Low
Organization	3.33	3.00	Low
Mechanics	3.00	3.00	Low
In-course Content	3.00	3.00	Low

This test paper demonstrates solid, consistent performance across all assessment criteria. Language use achieved 3.50, representing the strongest dimension, while vocabulary and organization both scored 3.33, indicating good competence. Mechanics and content appropriateness, though lower at 3.00, remain within acceptable parameters.

The low standard deviations throughout all categories constitute a particularly positive finding, suggesting that students performed with considerable uniformity. This consistency may reflect effective instruction or appropriate task design that allowed most learners to demonstrate their abilities adequately. The overall profile suggests a group of students operating at a similar proficiency level with no extreme outliers in either direction.

Table (30): Statistical Overview and Interpretation of Test Paper 6

Assessment Criterion	Mean Score	Median	Standard Deviation
Language	3.33	3.00	Low
Vocabulary	3.33	3.00	Low
Organization	3.17	3.00	Low
Mechanics	3.00	3.00	Low

In-course Content	3.00	3.00	Low

Test Paper 6 reveals satisfactory performance, though with modest room for improvement in several areas. Language and vocabulary both achieved 3.33, representing the strongest components of this assessment. Organization scored slightly lower at 3.17, while mechanics and content both reached 3.00.

The relatively lower scores in organization and mechanics suggest these areas may benefit from targeted instructional intervention. However, the low standard deviations indicate that performance was consistent across the sample, with students demonstrating similar competency levels. The overall pattern suggests adequate but not exceptional writing skills, with students meeting basic expectations without significantly exceeding them.

Table (31): Statistical Overview and Interpretation of Test Paper 7

Assessment Criterion	Mean Score	Median	Standard Deviation
Language	3.50	3.50	Low
Vocabulary	3.33	3.50	Low
Organization	2.83	3.00	Moderate
Mechanics	2.83	3.00	Moderate
In-course Content	2.83	3.00	Moderate

This test paper presents an interesting divergence in performance patterns. Language and vocabulary performed respectably at 3.50 and 3.33 respectively, suggesting students possessed adequate linguistic and lexical resources. However, organization, mechanics, and content all scored lower at 2.83, indicating challenges in these domains.

The moderate variability in median scores for the lower-performing categories suggests inconsistent application of organizational principles and mechanical conventions. This pattern may indicate that while students have developed reasonable language proficiency, they struggle to deploy that knowledge within well-structured, mechanically accurate texts. Such findings underscore the multidimensional nature of writing competence and the need for integrated skills development.

Table (32): Statistical Overview and Interpretation of Test Paper 8

Assessment Criterion	Mean Score	Median	Standard Deviation
Language	4.67	5.00	Very Low
Vocabulary	4.67	5.00	Very Low
Organization	5.00	5.00	Very Low
Mechanics	5.00	5.00	Very Low
In-course Content	4.83	5.00	Very Low

Test Paper 8 represents exceptional performance across all assessment dimensions, with scores ranging from 4.67 to 5.00. Organization and mechanics both achieved perfect scores of 5.00, indicating exemplary structural coherence and technical accuracy. Language, vocabulary, and content all scored 4.67 or higher, with median values reaching the maximum score of 5.00 in most categories.

The remarkably low standard deviations demonstrate that this outstanding performance was not limited to a few high-achieving individuals but rather characterized the entire sample. Such uniformly excellent results may reflect either a particularly strong cohort of students, an especially well-designed and appropriate task, or possibly both factors working in conjunction. This test paper serves as a benchmark for what students can achieve when performing at their optimal level.

Table (33): Statistical Overview and Interpretation of Test Paper 9

Assessment Criterion	Mean Score	Median	Standard Deviation
Language	3.00	3.00	Low
Vocabulary	3.50	3.50	Low
Organization	3.00	3.00	Low
Mechanics	2.83	3.00	Moderate
In-course Content	3.33	3.00	Moderate

This test paper demonstrates average to satisfactory performance across most dimensions. Vocabulary emerged as the strongest area at 3.50, followed by content at 3.33. Language and organization both achieved 3.00, representing adequate but

unremarkable competence. Mechanics scored lowest at 2.83, with moderate variability suggesting inconsistent application of grammatical and orthographic conventions.

The overall pattern indicates that students possess reasonable lexical resources and can generate relevant content, but struggle somewhat with mechanical accuracy and organizational coherence. The moderate standard deviations in mechanics and content suggest that performance varied more widely in these areas, with some students demonstrating better control than others. Such variability points to differential learning outcomes that may require attention in subsequent instruction.

Table (34): Statistical Overview and Interpretation of Test Paper 10

Assessment Criterion	Mean Score	Median	Standard Deviation
Language	4.67	5.00	Very Low
Vocabulary	4.67	5.00	Very Low
Organization	4.83	5.00	Very Low
Mechanics	4.50	4.50	Very Low
In-course Content	4.67	5.00	Very Low

Similar to Test Paper 8, this assessment reveals exceptional performance across all measured dimensions. Organization achieved the highest mean of 4.83, with language, vocabulary, and content all scoring 4.67. Even mechanics, often the weakest area in other assessments, reached a strong 4.50. The median values of 5.00 in most categories indicate that the majority of students performed at or near the maximum level.

The consistently very low standard deviations demonstrate remarkable homogeneity in performance, suggesting that excellence was widespread rather than concentrated among a few high achievers. This uniformly strong performance pattern may indicate significant growth in student competence or particularly effective task design that allowed learners to showcase their abilities optimally. The results establish a high standard that aligns with advanced proficiency expectations.

4.2.2 Summary of the Analytical Rubric and Writing Assessment Results

An analytical rubric was utilized to assess student writing performance, grounded in the idea that writing consists of multiple distinct components (Weigle, 2002). The rubric, adapted from Jacobs et al. (1981), Weir (1990),

and aligned with the Common European Framework of Reference for Languages (Council of Europe, 2002; 2009a, 2009b), evaluated five key aspects: content, organization, language use, vocabulary, and mechanics. Each dimension was scored on a scale from 0 (lowest) to 5 (highest), yielding a maximum total score of 25 points. Prior to implementation in this study, the rubric underwent piloting and refinement by two external EFL experts to ensure both reliability and validity.

The statistical analysis of writing test papers across ten different samples revealed distinct performance patterns across the assessed dimensions. Language proficiency scores generally ranged between 2.83 and 4.67, indicating mostly acceptable to very good command of English, with notable consistency observed across most samples. Vocabulary deployment demonstrated mean scores between 2.80 and 4.67, reflecting a satisfactory to strong lexical range utilized by students. Organizational competence varied considerably, with scores spanning from 2.83 to 5.00, suggesting generally good structural ability though with some variability depending on the assessment occasion. Mechanical accuracy, encompassing grammar, punctuation, and spelling, showed mean scores ranging from 2.33 to 5.00, typically falling within acceptable parameters, though certain cases exhibited inconsistency. Content relevance and appropriateness ranged from 2.83 to 4.83, reflecting generally appropriate and course-aligned writing across most student work.

Overall, the quality of writing performance was rated from satisfactory to very good across the evaluated dimensions. While numerous papers demonstrated strong and consistent performance, particularly in vocabulary usage and organizational structure, some variability emerged in language use and mechanical accuracy, pointing toward areas requiring targeted instructional intervention. This analysis was framed through a thematic analytical approach, integrating both quantitative statistical patterns and qualitative interpretation. As part of the broader investigation into writing assessment validity within Libyan higher education EFL contexts, document analysis was conducted on ten previous writing test papers to examine how teachers' assessment practices,

when applied to authentic student outputs, reflect established principles of writing validity through systematic rubric application.

4.2.3.1 Detailed Analysis of Performance Patterns Across Test Papers

The comprehensive analysis of all ten test papers reveals considerable variation in student performance across different assessment occasions, with several distinct patterns emerging that merit detailed discussion.

Exceptional Performance Group: First, Test Papers 8 and 10 stand out dramatically from the overall dataset, with scores consistently ranging between 4.50 and 5.00 across all five assessment dimensions. These two assessments demonstrate what students can achieve under optimal conditions, whether due to particularly effective preparation, appropriate task difficulty, favorable assessment circumstances, or a combination of these factors. The uniformly high performance across both papers, coupled with very low standard deviations, suggests these results represent genuine competence rather than isolated achievements. As such, these assessments may serve as important benchmarks for establishing program goals and illustrating target proficiency levels that the curriculum aims to develop systematically across all learners.

Strong Intermediate Performance Group: Second, Test Papers 3 and 5 exhibited good performance levels, with scores generally distributed between 3.17 and 4.00 across the various dimensions. Notably, these papers demonstrated consistent results with low standard deviations, suggesting stable competence at a solid intermediate level. This consistency indicates that students performing at this level have developed reliable writing skills that they can deploy with reasonable predictability. The performance pattern suggests internalization of writing conventions and strategies that enable consistent production of acceptable academic writing, though with room for advancement toward the excellence demonstrated in Papers 8 and 10.

Average to Satisfactory Performance Group: Third, Test Papers 1, 2, 6, 7, and 9 clustered around average to satisfactory performance levels, with scores typically ranging from 2.80 to 3.50 across the assessed dimensions. This represents the largest group of test papers, indicating that most assessment

occasions yielded adequate but clearly improvable writing performance. Students in this performance band demonstrate basic competence in EFL academic writing but have not yet achieved the consistency or sophistication evident in higher-performing papers. The clustering of multiple test papers within this range suggests this may represent a typical or expected performance level for students at this stage of their language development, highlighting the need for continued instructional support to move learners beyond adequate performance toward stronger proficiency.

Problematic Performance Case: Test Paper 4 presented unique challenges and warrants separate discussion. This assessment revealed particular difficulties regarding language use and overall consistency, as evidenced by relatively higher standard deviations compared to other test papers. The mean language score of 2.83 with a notably lower median of 2.50 indicates that a substantial portion of students struggled significantly with linguistic accuracy and appropriateness on this particular assessment. This variability suggests either that the task itself posed unusual difficulty that affected students differentially based on their proficiency levels, or that the student sample for this particular test exhibited heterogeneous proficiency levels that were not apparent in other assessments. This finding warrants further investigation to determine whether task characteristics, administration conditions, or sample composition factors contributed to the observed inconsistency.

Cross-Cutting Patterns (Mechanics as a Persistent Challenge):

Examining performance patterns across all assessments reveals that mechanics consistently emerged as a challenging area, with several test papers identifying this as the weakest dimension. Even in otherwise strong performances, mechanical accuracy often lagged behind other aspects of writing quality. This persistent pattern suggests a systemic need for enhanced attention to grammatical accuracy, punctuation conventions, and spelling in the instructional program. The consistency of this weakness across multiple assessment occasions indicates that current instructional approaches may not adequately address mechanical aspects of writing, or that students require more extensive practice and feedback in this domain before achieving automaticity

in applying mechanical conventions accurately.

Cross-Cutting Patterns (Variability in Organization): Organization demonstrated more variable performance across the test papers, with scores ranging from weak (2.83) to excellent (5.00). This substantial variation is noteworthy because it suggests that students' ability to structure coherent, well-organized texts may depend significantly on task characteristics, genre familiarity, topic knowledge, or other contextual factors. Unlike vocabulary or language use, which tend to reflect relatively stable underlying competencies, organizational skill appears more susceptible to situational variables. This finding has important implications for instruction, suggesting that students may benefit from explicit teaching of organizational strategies across different text types and writing contexts, rather than assuming that organizational competence will transfer automatically from one writing situation to another.

Pedagogical and Assessment Implications: These findings underscore the complex, multifaceted nature of writing competence in EFL contexts. Writing ability does not develop uniformly across all dimensions; rather, students may demonstrate relative strengths in certain aspects (such as vocabulary or content) while continuing to struggle with others (particularly mechanics). The analytical rubric employed in this study proved valuable precisely because it enabled identification of these specific patterns, allowing for targeted diagnosis of areas requiring development within the EFL writing curriculum at Sabratha University's Faculty of Education.

Moreover, the substantial variation in performance across different test papers raises important questions about assessment consistency and validity. The fact that some assessments yielded predominantly excellent performance while others resulted in merely satisfactory outcomes suggests the need to examine whether tasks are appropriately calibrated for difficulty and whether assessment conditions remain sufficiently consistent across different administration occasions. Future research and program development efforts should address these concerns to ensure that writing assessments provide reliable, valid indicators of student competence that can inform both instructional decisions and student progress evaluation.

4.3 Thematic Analysis Framework

The document analysis was conducted using thematic analysis, allowing for the identification of recurring patterns and themes across multiple exam samples. These themes were interpreted through both statistical outcomes and qualitative criteria derived from an analytical scoring rubric covering five key components: language, vocabulary, organization, mechanics, and content relevance (in-course content).

Theme 1: Emphasis on Multi-Dimensional Assessment

The use of an analytical rubric reflects a clear shift towards a multi-dimensional evaluation of writing, supporting the notion of construct validity. Teachers were guided to assess distinct elements of student writing, acknowledging that writing proficiency involves more than grammatical correctness alone. This supports findings by Weigle (2002) and Hamp-Lyons (1991) who advocate for analytical rubrics in EFL contexts due to their diagnostic and pedagogical benefits.

Theme 2: Consistency and Reliability in Assessment Practices

Analysis of the ten writing test papers revealed notable consistency in mean and median scores, particularly across the domains of language, vocabulary, and organization, demonstrating a high degree of scoring reliability. The close alignment between median and mean values observed in many papers suggests that teacher ratings were both stable and replicable. This pattern indicates positive progress toward achieving inter-rater reliability and internal consistency, both of which constitute essential components of valid assessment practice.

Theme 3: Evident Construct Validity through Task-Relevant Evaluation

The analytical rubric's five-dimensional structure—language, vocabulary, organization, mechanics, and content relevance—demonstrates strong construct validity, substantiated by triangulated evidence from test paper analysis, teacher questionnaires, and interviews.

Multi-Dimensional Assessment in Practice: Teacher questionnaire responses revealed systematic attention to multiple writing dimensions. When asked "To what extent do you consider content relevance when assessing student writing?" some of teachers reported "always" or "frequently" evaluating content alongside linguistic accuracy. While, others indicated they "distinguish between different aspects of writing when scoring" (e.g., organization vs. grammar), with many attributing this differentiation to rubric use. As one teacher explained: "The rubric helps me see organization separately from grammar—sometimes students organize well but make many errors, and the rubric lets me give credit for both aspects fairly". Interview data corroborated these patterns. When asked "In your experience, can students have good grammar but poor content, or vice versa?" teachers consistently recognized dimensional independence. One teacher observed, "I've had students who write beautifully correct sentences but don't really say anything meaningful. And I've had others who have great ideas but struggle to express them without errors. That's why we need to assess both separately". Responses to "Can you describe what you look for when evaluating the 'content' aspect?" revealed sophisticated criteria extending beyond surface features. Teachers emphasized assessing whether students "understand the main concepts" and "address the task properly with relevant information," rather than focusing solely on linguistic accuracy.

Empirical Evidence from Test Papers: Quantitative analysis provided concrete evidence of dimensional differentiation in actual scoring practices. Test Paper 9 demonstrated vocabulary scores (3.50) exceeding language scores (3.00), indicating teachers distinguished between lexical resources and overall proficiency. Test Paper 4 showed even starker divergence: organization (3.67) rated substantially higher than language use (2.83), demonstrating that teachers recognized and rewarded structural coherence despite linguistic weaknesses. Test Paper 7 exhibited the inverse pattern, with language (3.50) and vocabulary (3.33) exceeding organization, mechanics, and content (all 2.83), revealing students who possessed linguistic resources but struggled to deploy them effectively. Content relevance scores ranged from 2.83 to 4.83 across all papers, varying substantially and independently from linguistic dimensions. Test Papers 8 and 10 achieved content scores of 4.83 and 4.67 respectively, demonstrating excellence in both linguistic accuracy and substantive communication. Conversely, papers scoring around 3.00 for content showed that adequate linguistic skills do not

automatically translate to effective academic communication. These scoring patterns would be impossible if teachers simply formed global impressions and assigned similar scores across categories. The documented dimensional independence demonstrates genuine differentiation in evaluation practices.

Construct Validity and Theoretical Alignment: This comprehensive approach represents a departure from purely form-focused assessment traditions. By requiring evaluation of whether students successfully communicate course-relevant academic content, the rubric operationalizes an authentic conception of writing ability aligned with contemporary sociocultural and genre-based theories. Writing is assessed as purposeful, context-bound activity rather than merely grammatical knowledge display. The convergence of evidence—statistical patterns showing dimensional independence, questionnaire responses indicating systematic multi-criteria attention, and interview narratives describing sophisticated evaluation distinguishing form from function—demonstrates that teachers have operationalized a theoretically sound conception of EFL academic writing competence. This conception extends beyond surface features to encompass communicative competence essential for academic success, thereby strengthening construct validity and the evidentiary basis for inferences about students' actual writing abilities and instructional needs (Messick, 1989).

Theme 4: Recognition of Variability and Diagnostic Feedback

While overall scores were generally within acceptable to high ranges, variability in mechanics and organization scores highlighted areas where students needed further development. This theme suggests that teachers used the rubric not only for summative purposes but potentially as a tool for formative, diagnostic assessment. For example, standard deviation values revealed inconsistency in grammar and spelling, pointing to common areas of difficulty in EFL writing instruction.

Theme 5: Differentiated Judgments Reflecting Student Performance Levels

Several papers, particularly those with higher scores, displayed consistently strong performance across all five dimensions, with mean scores close to the maximum of 5.0. Conversely, mid- range papers exhibited fluctuating scores, especially in mechanics and organization. This variation reflects a

differentiated judgment approach, where teachers recognize varying levels of performance rather than applying a holistic or norm-referenced method. Such an approach supports scoring validity and aligns with fair assessment practices.

4.4 Summary of the Chapter

This chapter presented a mixed-methods analysis of EFL writing assessment validity at Sabratha University, examining ten test papers scored by six experienced teachers using an analytical rubric alongside teacher and student questionnaires. The rubric assessed five dimensions—language, vocabulary, organization, mechanics, and content relevance—each scored 0-5. Teachers administered assessments regularly (83% weekly or monthly) to assess language proficiency, employing rubrics and established standards equally to ensure validity. Students showed moderate confidence (50% "somewhat confident"), limited practice (15% daily), and identified vocabulary (55%) and grammar (30%) as improvement areas. Performance analysis revealed exceptional scores in Papers 8 and 10 (4.50-5.00), strong performance in Papers 3 and 5 (3.17-4.00), and average performance in remaining papers (2.80-3.50). Critical patterns showed mechanics as consistently weakest, suggesting instructional gaps, while organization varied substantially (2.83-5.00), indicating task-dependence. Findings demonstrate theoretically sound assessment practices with the rubric successfully identifying strengths and weaknesses, though areas requiring attention include persistent mechanics weakness, organizational competence variability, limited student practice, and task calibration consistency warranting future investigation.

Chapter Five: Discussion

5.0 Introduction

This chapter presents a discussion of the findings from the analysis of both the teacher and student questionnaires, along with the document analysis of previous writing exams. It critically evaluates the results in light of the research questions and existing literature on the validity of writing assessments in EFL contexts. The primary aim of this chapter is to provide an interpretation of the findings, compare them to prior studies, and explore how they contribute to the understanding of writing assessment validity in Libyan higher education. The chapter is organized around the key research questions and synthesizes the findings from the different research tools used in the study.

5.1 Research Question One

Q1. Do EFL learners writing skills accurately assessed using the right EFL writing assessment tools?

The findings indicate that EFL learners' writing skills are assessed with reasonable accuracy through the analytical rubric employed, though certain limitations exist. The five-dimensional rubric evaluating language, vocabulary, organization, mechanics, and content relevance effectively captured distinct competencies, as evidenced by teachers' ability to differentiate among dimensions—vocabulary scores exceeded language proficiency in some papers while organizational competence surpassed both in others. Teachers' questionnaire responses confirmed systematic attention to multiple criteria, with half employing rubrics for specific aspects and half comparing performance against established standards. However, accuracy faces three challenges: mechanics consistently emerged as the weakest dimension, suggesting genuine student difficulty or inconsistent application of criteria; organizational scores varied dramatically from 2.83 to 5.00 depending on tasks, questioning whether assessments measure stable competence or task-specific performance; and exceptional results in Papers 8 and 10 contrasted sharply with average performance elsewhere, raising concerns about task calibration consistency. While the analytical rubric represents a theoretically sound tool aligned with construct validity principles, its effectiveness depends on consistent application, appropriate task design, and teachers continued professional development.

5.2 Research Question Two

Q2. What procedures and techniques do EFL teachers use to ensure the validity of writing assessments?

EFL teachers employ multiple procedures to establish assessment validity, revealing both strengths and areas requiring development. For content validity, half connect prompts with curriculum standards while a third incorporate real-world contexts. For construct validity, teachers divide equally between comparing performance against established standards and employing analytical rubrics that disaggregate writing dimensions. Teachers address reliability through clear scoring criteria (50%) and multiple raters (33%), though few offer revision opportunities. The analytical rubric serves as a central validity mechanism, requiring independent evaluation of five dimensions and reducing global impression bias. However, teachers acknowledge persistent challenges, particularly balancing accuracy against fluency (50%) and differentiating developmental errors from competence gaps (33%). Teachers' feedback practices—predominantly peer review and conferences—indicate formative orientations supporting consequential validity, with half using results to identify individual needs. Nevertheless, substantial performance variation across papers and persistent mechanics weakness suggests that while teachers employ appropriate validity-supporting procedures conceptually, implementation consistency and task calibration require ongoing refinement.

5.3 Discussion of the Questionnaire Findings.

Both sources, teacher and student questionnaire, confirmed that writing is frequently assessed in classrooms with teachers using regular assessments to evaluate student proficiency. The student questionnaires revealed limited out-of-class writing engagement. This discrepancy suggests that despite frequent assessments, students may not be engaging in sustained, authentic writing practice outside the classroom, which could affect the ecological and consequential validity of the assessments. These finding aligns with Weigle (2002), who emphasizes that frequent assessments must be paired with consistent writing practice to ensure ecological and consequential validity. Without sustained practice outside the classroom, students may not have the opportunity to apply writing skills in authentic contexts, which undermines the validity of assessments. This

finding highlights the need for assessments to not only to be frequent but also aligned with real-world writing experiences.

Both teachers and students recognized grammar, structure, and clarity as recurring challenges in writing. These difficulties were evident in the rubric analysis, where variability was observed in mechanics and organization scores. This shared recognition enhances the face and construct validity of the assessment tools, suggesting that the rubric captures the core challenges faced by students. This mirrors the findings from Hamp-Lyons (1991) and Leki (1990), who identified similar structural and grammatical challenges in EFL writing. The shared acknowledgment of these issues suggests that the assessments are targeting the right areas but also indicates a need for more targeted instructional support in these areas.

5.4 Rubric Use and Reliability and Student Awareness Gap

Teachers reported using rubrics and multiple raters to support the reliability of assessments, which was confirmed by the consistent pattern of scores in the rubric-based analysis. However, students appeared largely unaware of the rubric criteria, indicating a gap in assessment literacy. This lack of awareness could contribute to students overestimating their writing abilities, thus limiting the formative potential of assessments. These finding aligns with Hamp-Lyons (1991), who highlighted the importance of assessment transparency. When students do not understand the rubric, they may misinterpret feedback and overestimate their abilities, leading to misaligned self-assessment. This underscores the need for greater communication between teachers and students regarding assessment criteria.

5.5 Feedback and Practices

Both teachers and students acknowledged the importance of feedback. Teachers emphasized strategies such as peer review and individual conferences, and students reported frequent revisions. However, the rubric analysis revealed that feedback often resulted in surface-level rather than substantive revisions, suggesting that students were not fully engaging with the feedback provided. This similar to the finding of Leki (1990), who observed that while students receive feedback, the quality and depth of revisions often remain superficial. Previous studies, such as Weigle (2002), emphasize that feedback should be actionable and guide students in making meaningful revisions. Strengthening the feedback loop to ensure deeper engagement with feedback could

enhance the validity of writing assessments by promoting real improvement in students' writing skills.

5.6 Formal Language as a Common Challenge

Both the rubric analysis and the questionnaire findings identified formal academic language as a challenge for many students. The rubric scores showed some inconsistency in language use, while students reported feeling uncomfortable using formal academic language. This convergence suggests that issues with formal language are not isolated but are widespread among students. The finding is consistent with Weigle (2002) and Leki (1990), who noted that many EFL students struggle with formal language, which is a key component of content validity in writing assessments. The difficulty students face with academic language underscores the need to include formal language as an explicit criterion in writing rubrics to enhance construct validity.

5.7 Assessment-Instruction Link

The rubric-based analysis indicated that assessments were used to inform instruction, with teachers adapting their teaching based on assessment outcomes. However, students' moderate engagement with feedback and revision suggests a gap between instructional intentions and students' application of feedback, highlighting a disconnect between assessment use and student learning. This divergence echoes findings from Hamp-Lyons (1991), where teachers used assessments to guide instruction, but students did not always engage fully with feedback. This points to the need for instructional strategies that not only provide feedback but also guide students on how to apply it effectively to improve their writing.

5.8 The need for Improvement

Teachers recommended incorporating more authentic tasks and exposing students to diverse writing genres, which aligned with weaknesses identified in the rubric-scored essays and the students' reported struggles with essay writing and genre familiarity. While students did not explicitly request changes, their challenges with writing reinforced teachers' calls for reform. Weigle (2002), who advocated for genre-based writing tasks in EFL contexts to better prepare students for real-world writing situations. The call for authentic tasks reflects the need for assessments to be

ecologically valid, ensuring that they reflect the types of writing tasks students will encounter in their academic and professional lives.

5.9 Summary of the Chapter

The comparison of the rubric-based document analysis with the questionnaire findings reveals significant areas of convergence, including shared recognition of writing difficulties, the importance of feedback, and challenges with formal language. However, discrepancies regarding student self-assessment, the application of feedback, and the lack of out-of-class writing practice suggest several areas for improvement. These findings emphasize the need for transparent, criterion-based assessments, authentic tasks, and better integration of feedback to improve the validity and reliability of writing assessments in Libyan EFL higher education. Addressing these gaps will enhance both instructional practices and student learning outcomes, ultimately contributing to more valid and effective writing assessments in the context.

Chapter Six: Conclusion

6.0 Introduction

The study aimed to assess the validity of writing assessments in Libyan EFL higher education by examining the perceptions of both teachers and students, alongside an analysis of previous writing exams. This final chapter concludes the study by synthesizing the key findings, discussing their implications, and providing recommendations for various stakeholders. It also acknowledges the limitations of the research and suggests directions for future inquiry.

6.1 Conclusion of the whole study

This study concludes that while writing is frequently assessed in Libyan EFL higher education, a significant gap exists between classroom assessment and real-world writing application. A notable discrepancy was found in students' out-of-class writing practices, which impacts the ecological and consequential validity of current assessments. Both teachers and students identified common challenges in grammar, structure, and clarity, confirming that the assessment tools are targeting relevant areas of writing competence.

However, a critical issue emerged regarding assessment literacy; while teachers consistently used rubrics to ensure reliability, students were largely unaware of these criteria. This leads to a potential overestimation of their abilities and hinders the formative value of feedback. The study also found that feedback practices, though acknowledged as important, often resulted in superficial revisions rather than substantive improvements. Furthermore, both rubric scores and student self-reports highlighted a shared struggle with formal academic language. Finally, a disconnect between the instructional use of assessments by teachers and students' limited engagement with feedback suggests a need for reform, with teachers calling for more authentic and diverse writing tasks to better align assessment with learning.

6.2 Implication of the Study

The findings of this study have several important implications for educators, policymakers, and curriculum designers in Libyan EFL higher education. The identified gap in assessment literacy implies an urgent need for more transparent assessment practices. If students do not understand how their writing is evaluated, the formative

potential of assessment is lost. Therefore, institutions and teachers should prioritize making assessment criteria explicit.

The study also implies that current feedback strategies may not be effective in promoting deep learning. The tendency for students to make only superficial revisions suggests that feedback must be more actionable and better integrated into the teaching and learning cycle. Furthermore, the call for more authentic writing tasks implies that current assessments may lack ecological validity, failing to prepare students for the academic and professional writing demands they will face. This points to a need for curriculum reform that incorporates tasks mirroring real-world writing challenges, thereby bridging the gap between classroom assessment and practical writing competence.

6.3 Recommendation of the Study

Based on the results of this study, the following are recommended:

6.3.1 Recommendation for Teachers

Based on the study's findings, the following recommendations are offered to EFL teachers:

- Teachers should explicitly share and discuss assessment rubrics with students before assignments are given. This will help bridge the awareness gap, manage student expectations, and empower them to self-assess their work more accurately.
- Feedback should be focused and constructive, guiding students toward substantive revision rather than just surface-level error correction. Incorporating peer feedback and one-on-one conferences can help ensure students understand and can apply the feedback effectively.
- Teachers are encouraged to design assessments that include authentic, real-world writing tasks and a wider variety of genres. This will improve the ecological validity of assessments and better prepare students for future academic and professional contexts.
- To close the disconnect between assessment and learning, teachers should foster an environment of open communication, discussing the purpose of assessments and how feedback can be used as a tool for improvement.

6.3.2 Recommendation for Students

To improve their writing competence, students are encouraged to:

- Engage in regular writing practice outside of formal class assignments to build fluency and confidence. This can include journaling, blogging, or participating in online writing forums.
- Students should take the initiative to understand the rubrics and criteria by which their writing is assessed. Asking teachers for clarification can help turn assessment from a mere grade into a valuable learning tool.
- View feedback not as criticism, but as guidance for improvement. Students should move beyond correcting only grammatical errors and use comments to rethink the structure, clarity, and depth of their ideas.
- By understanding the assessment criteria, students can learn to evaluate their own writing, identify areas for improvement, and become more autonomous learners.

6.4 Limitation of the Study

While this study provides valuable insights, it is important to acknowledge several limitations.

- **Contextual Limitation:** The study was conducted in a specific cultural and educational context, which may limit the generalizability of the findings to other EFL settings.
- **Sample Size:** The study relied on a limited number of teachers and students, which may not fully represent the diverse experiences and perspectives of all EFL learners and instructors in Libyan higher education.
- **Focus on Writing Assessments:** The study concentrated solely on writing assessments, which may exclude other important aspects of EFL assessment practices, such as speaking, listening, and reading.

6.5 Suggestions for further Research

Based on the findings of this study, several directions for future research are recommended:

- Conduct longitudinal studies to track the effectiveness of revised assessment practices over time and evaluate their long-term impact on student writing development.
- Examine the validity of EFL writing assessments in different cultural and educational environments to determine whether similar challenges exist across various regions.
- Compare various educational levels, such as secondary versus higher education, to provide insights into how assessment practices evolve at different stages of learning.
- Explore the factors that influence how students interpret and apply feedback to improve their writing skills.
- Investigate how professional development and teacher training impact the quality of writing assessment practices.
- Examine how digital tools and online platforms can support more valid and efficient writing evaluation processes.

6.6 Summary of the Thesis

This study has provided a comprehensive analysis of writing assessment practices in Libyan EFL higher education, shedding light on areas of strength and areas needing improvement. While the frequent assessment of writing and recognition of common writing challenges reflect positive aspects of the current system, issues such as limited student engagement with writing practice, feedback, and rubrics highlight the need for significant reform. By addressing these gaps, Libyan EFL education can move closer to achieving assessments that truly reflect students' writing competencies and enhance their learning outcomes. Ultimately, the study reinforces the importance of aligning assessment tools, instructional practices, and student engagement to improve the validity and reliability of writing assessments in EFL contexts.

:

References

1. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
2. Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall.
3. Andrade, H. G. (2000). Using rubrics to promote thinking and learning. Educational *Leadership, 57*(5), 13-18.
4. Andrade, H. G. (2001). The effects of instructional rubrics on learning to write. *Current Issues in Education, 4*(4), 1-22.
5. Andrade, H. G., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research & Evaluation, 10*(3), 1-11.
6. Andrade, H. L., & Cizek, G. J. (2010). *Handbook of formative assessment.* Routledge.
7. Arter, J., & McTighe, J. (2001). Scoring rubrics in the classroom: Using performance *criteria for assessing and improving student performance*. Corwin Press.
8. Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
9. Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.
10. Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*(1), 31-57.
11. Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74.
12. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101.
13. Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.
14. Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Utah State University Press.
15. Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist, 32*(7), 513-531.

16. Brookhart, S. M. (2013). How to create and use rubrics for formative assessment and *grading*. ASCD.
17. Brown, H. D. (1994). *Principles of language learning and teaching* (3rd ed.). Prentice Hall.
18. Brown, H. D. (2003). *Language assessment: Principles and classroom practices.* Longman.
19. Brown, H. D. (2004). *Language assessment: Principles and classroom practices.* Pearson Education.
20. Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.
21. Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Rand McNally.
22. Choi, I. C., Lee, Y. J., & Kang, S. J. (2009). Implementing a task-based assessment in a Korean EFL context. *Language Assessment Quarterly, 6*(4), 342-358.
23. Connor, U. (1996). *Contrastive rhetoric: Cross-cultural aspects of second-language writing.* Cambridge University Press.
24. Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Lawrence Erlbaum.
25. Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.
26. Crusan, D. (2010). Assessment in the second language writing classroom. University of Michigan Press. https://doi.org/10.3998/mpub.770334
27. Cumming, A. (1997). The testing of second language writing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (pp. 51-63). Kluwer Academic.
28. Cumming, A. (2001). Learning to write in a second language: Two decades of research. *International Journal of English Studies, 1*(2), 1-23. https://doi.org/10.6018/ijes.1.2.48331
29. Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing Writing, 8*(2), 73-83.
30. Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67-96.

31. Davies, A. (1968). *Language testing symposium: A psycholinguistic approach*. Oxford University Press.
32. Dobric, V. (2018). Validity in educational assessment: A systematic review of the literature. *Educational Research Review, 24*, 52-71.
33. Dorans, N. J., & Cook, L. L. (Eds.). (2016). *Fairness in educational assessment and measurement*. Routledge.
34. Dweni, D. K. (2023). Challenges encountered by Libyan EFL undergraduate students in English research writing. Scientific Journal of Faculty of Education, Misurata University, 9(21), 312
35. Elbow, P., & Belanoff, P. (1997). Reflections on an explosion: Portfolios in the '90s and beyond. In K. B. Yancey & I. Weiser (Eds.), *Situating portfolios: Four perspectives* (pp. 21-33). Utah State University Press.
36. Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.
37. Eswaey, G., & Ihmoumah, H. (2024). Role of Self-Assessment in Improving Students' Writing: A Systematic Review. AlQalam Journal of Medical and Applied Sciences, 7(Supp 2), 94–106. https://doi.org/10.54361/ajmas.2472214
38. Ferris, D. R. (2011). *Treatment of error in second language student writing* (2nd ed.). University of Michigan Press.
39. Ferris, D. R., & Hedgcock, J. S. (2014). *Teaching L2 composition: Purpose, process, and practice* (3rd ed.). Routledge.
40. Folse, K. S. (2004). *Vocabulary myths: Applying second language research to classroom teaching.* University of Michigan Press.
41. Folse, K. S., Muchmore-Vokoun, A., & Solomon, E. V. (2010). *Great writing 4: Great essays* (3rd ed.). Heinle Cengage Learning.
42. Fox, J. (2004). Test decisions over time: Tracking validity. *Language Testing, 21*(4), 437-465.
43. Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.
44. Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective.* Longman.
45. Hacker, D., & Sommers, N. (2011). *A writer's reference* (7th ed.). Bedford/St. Martin's.

46. Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.
47. Hamp-Lyons, L. (1990). *Second language writing: Assessment issues*. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69–87). Cambridge University Press.
48. Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Ablex.
49. Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 162-189). Cambridge University Press.
50. Hamp-Lyons, L., & Condon, W. (2000). *Assessing the portfolio: Principles for practice, theory, and research*. Hampton Press.
51. Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice, 4*(3), 365-379.
52. Hillocks, G. (2002). *The testing trap: How state writing assessments control learning*. Teachers College Press.
53. Holden, R. R. (2010). Face validity. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini encyclopedia of psychology* (4th ed., Vol. 2, pp. 637-638). John Wiley & Sons.
54. Huang, J., Shear, L., & Stevenson, H. (2016). The impact of test design on student engagement with test tasks. *Educational Assessment, 21*(2), 83-101.
55. Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*(2), 219-230.
56. Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press.
57. Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47*(4), 549-566.
58. Huot, B. (2002). *Rearticulating writing assessment for teaching and learning*. Utah State University Press.
59. Huot, B., & Neal, M. (2006). Writing assessment: A techno-history. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 417-432). Guilford Press.

60. Hyland, K. (2003). *Second language writing*. Cambridge University Press.
61. Hyland, K. (2009). *Teaching and researching writing* (2nd ed.). Pearson Education.
62. Irons, A. (2008). *Enhancing learning through formative assessment and feedback.* Routledge.
63. Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Newbury House.
64. Johns, A. M. (1997). *Text, role, and context: Developing academic literacies.* Cambridge University Press.
65. Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144.
66. Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). American Council on Education/Praeger.
67. Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.
68. Kohn, A. (2006). The trouble with rubrics. *English Journal, 95*(4), 12-15.
69. Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1-14). Cambridge University Press.
70. Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. Longraw.
71. Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387-431). American Council on Education/Praeger.
72. Langan, J. (2010). *College writing skills with readings* (8th ed.). McGraw-Hill.
73. Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing, 17*(1), 43-64.
74. Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.
75. Lumley, T. (2005). Assessing second language writing: The rater's perspective. Peter Lang.
76. Lunsford, A. A., & Lunsford, K. J. (2008). "Mistakes are a fact of life": A national comparative study. *College Composition and Communication, 59*(4), 781-806.

77. McMillan, J. H. (2014). Classroom assessment: Principles and practice for effective standards-based instruction (6th ed.). Pearson.
78. McNamara, T. F. (1996). *Measuring second language performance*. Longman.
79. McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian Citizenship Test. *Language Assessment Quarterly, 8*(2), 161-178.
80. Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). American Council on Education/Macmillan.
81. Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.
82. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.
83. Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7*(2), 191-205.
84. Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research & Evaluation, 7*(3), 1-6.
85. Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10), 1-6.
86. Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge University Press.
87. Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement, 22*(4), 287-293.
88. Nitko, A. J., & Brookhart, S. M. (2007). *Educational assessment of students* (5th ed.). Pearson.
89. Oshima, A., & Hogue, A. (2006). *Writing academic English* (4th ed.). Pearson Education.
90. Oshima, A., & Hogue, A. (2007). *Introduction to academic writing* (3rd ed.). Pearson Education.
91. Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129-144.
92. Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

93. Popham, W. J. (2008). *Transformative assessment*. ASCD.
94. Ramadan, M. O., & Dekheel, H. (2020). Libyan students' perceptions of traditional exams as an assessment method: An exploratory study in the English Department at Sirte University. Abhat Journal, 16, 317–339. https://doi.org/10.37375/abhat.vi16.481
95. Read, J. (2000). *Assessing vocabulary.* Cambridge University Press.
96. Reid, J. M. (2000). *The process of composition* (3rd ed.). Longman.
97. Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research & Evaluation, 11*(10), 1-13.
98. Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159-179.
99. Schmitt, N. (2000). *Vocabulary in language teaching.* Cambridge University Press.
100. Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. *Infancy, 2*(4), 419-436.
101. Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39-83). Rand McNally.
102. Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8, 13, 24.
103. Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457-490.
104. Smalley, R. L., Ruetten, M. K., & Kozyrev, J. R. (2012). *Refining composition skills: Academic writing and grammar* (6th ed.). Heinle Cengage Learning.
105. Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher, 37*(4), 189-199.
106. Solórzano, R. W. (2008). High stakes testing: Issues, implications, and remedies for English language learners. *Review of Educational Research, 78*(2), 260-329.
107. Spandel, V. (2012). *Creating writers through 6-trait writing assessment and instruction* (6th ed.). Pearson.
108. Stiggins, R. J. (2001). *Student-involved classroom assessment* (3rd ed.). Prentice Hall.

109. Taras, M. (2005). Assessment – summative and formative – some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466-478.
110. Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). University of Minnesota, National Center on Educational Outcomes.
111. Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation, 9*(2), 1-7.
112. Topping, K. J. (2009). Peer assessment. *Theory into Practice, 48*(1), 20-27.
113. Truss, L. (2003). *Eats, shoots & leaves: The zero-tolerance approach to punctuation.* Gotham Books.
114. Valdés, G. (2001). Learning and not learning English: Latino students in American schools. Teachers College Press.
115. Valdés, G., Bunch, G., Snow, C., & Lee, C. (2005). Enhancing the development of students' language(s). In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world* (pp. 126-168). Jossey-Bass.
116. Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*. Cambridge University Press.
117. Waragh, I. M. S. (2016). Assessment Methods and Factors Affecting their Use by Libyan Tutors in Assessing Students' Writing and How These Assessment Methods Are Perceived by Students (PhD thesis, University of Sunderland).
118. Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.
119. Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.
120. White, E. M. (1985). *Teaching and assessing writing*. Jossey-Bass.
121. White, R., & Arndt, V. (1991). *Process writing.* Longman.
122. Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. Jossey-Bass.
123. Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Lawrence Erlbaum.
124. Wilson, M. (2006). Rethinking rubrics in writing assessment. *Journal of Writing Assessment, 2*(1), 5-22.

125. Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147-170.
126. Zieky, M. (2006). Fairness review in assessment. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359-376). Lawrence Erlbaum.
127. Zwick, R. (Ed.). (2006). *Rethinking the SAT: The future of standardized testing in university admissions*. Routledge Falmer.

APPENDIXES

Appendix (1) Teachers' Questionnaire

Greetings, Respondents

This survey was created as a means of gathering information to look into the teachers' opinions on the evaluating EFL learners' writing competence at the faculty of education at Sabratha University.

Kindly respond to every question in this survey. We are only interested in your own viewpoint; This questionnaire's answers are all kept private and confidential, and they will only be utilized for the purpose of the study. I appreciate your cooperation. Please provide your personal information of each statement by making a tick, / on the box.

1- gender: male □ female□

2- **Years of experience as an English language teacher:**

1-4□ 5-9 □ 10-14 □ 15-19 □ 20-24 □

3- Your age:

20-25□ 26-30□ 31 -34□ 35 -40□ 41-45□ 46-50□

4- **What is the average number of students in the classroom?**

5- **10-15□ 16-20□ 21-25□ 25-30□ 31-35□ 36-40□**

6- **level of education:**

Certificate. □ Diploma. □ Bachelor. □ Master. □ Doctorate. □

1- How frequently do you administer writing assessments in your EFL classroom?

a-Rarely □ b-Occasionally □

c - Monthly □ d-Weekly□ e-Other (please specify) … ……

2- How do you define the purpose of EFL writing assessments in your classroom?

a- Assessing language proficiency □

b- Promoting language development □

c- Evaluating students' understanding of writing conventions □

d- Other (please specify) ……..

3- What types of writing prompts do you typically use in EFL writing assessments?

a- Opinion/Argumentative □ b-Descriptive □

c-Narrative □ d-Expository □

e-Other (please specify) …….

4- How do you ensure that the writing prompts used in assessments are relevant and meaningful to your EFL students?

a- Aligning prompts with students' interests and experiences. □

b- Incorporating real-world contexts. □

c- Connecting prompts to the curriculum. □

d- Other (please specify) ……

5- How do you assess the validity of EFL writing assessments in your classroom?

a- Comparing students' performance with established writing standards. □

b- Using rubrics to evaluate specific writing criteria (e.g., grammar, vocabulary, organization). □

c- Analyzing the correlation between writing scores and other language proficiency measures (e.g., speaking, listening). □

d- Other (please specify) ……...…….

6-How do you address the issue of reliability in EFL writing assessments?

a- Using multiple raters to evaluate students' writing. □

b- Providing clear and consistent scoring criteria. □

c- Offering opportunities for students to revise and improve their writing. □

d- Other (please specify) ……

7-In your opinion, what are the biggest challenges in assessing EFL writing validity?

e- Differentiating between language errors and developmental stages□

f- Capturing the complexity of language use in a single assessment. □

g- Balancing the need for accuracy and fluency. □

h- Other (please specify) …….

8- How do you provide feedback to students based on their writing assessments?

a- Written comments on specific strengths and areas for improvement. □

b- One-on-one conferences to discuss their writing. □

c- Peer feedback and revision activities. □

d- Other (please specify) …….

9- How do you use the results of EFL writing assessments to inform your instruction?

a- Identifying individual student needs and designing targeted instruction. □

b- Adjusting the pace and content of the curriculum. □

c- Providing additional support or enrichment opportunities. □

d- Other (please specify) ……

10- What changes or improvements would you suggest to enhance the validity of EFL writing assessments?

a- Including a wider range of writing genres and formats. □

b- Incorporating authentic writing tasks. □

c- Providing more opportunities for students to engage in the writing process (pre-writing, drafting, revising) □.

d- Other (please specify) ……

Please provide any additional comments or insights you have regarding the validity of EFL writing assessments:

……

…….....……..

Thank you for participating in this questionnaire! Your feedback is greatly appreciated.

Appendix (2) Students' Questionnaire

Dear Candidates

This structured questionnaire is designed to EFL undergraduate students. The survey was created as a means of gathering information to look into the students' opinions on the evaluating EFL learners' writing competence at the faculty of education at Sabratha University. I would be grateful if you kindly complete this questionnaire. Your participation would help with the success of the study. The data collected through this survey will be kept strictly confidential and it will be used for the academic research only.

❖ Participants Personal Information:

Name \ Optional: .....

Gender: Male: ( ) Female: ( )

Age: ( )

University: …… College:

…… # Student's Background

Knowledge Challenges

- Please tick the right answer for each of the following questions:

1- How confident do you feel about your writing skills in English?

a-Very confident. □ b-Somewhat confident. □

c-Not very confident. □ d-Not confident at all. □

2- How often do you practice writing in English outside of the classroom?

a-Daily. □ b-Several times a week. □ c-Once a week. □

d-Rarely. □ e-Never. □

3- Which types of writing tasks do you find most challenging? (Select all -that apply)

a-Essay writing. □ b-Letter writing. □ c-Creative writing. □ d-Report writing. □

e- Other (please specify) ……

4- On a scale of 1-5, how well do you understand the structure and organization of a well-written paragraph or essay?

1- (Not at all). □ 2- little □ 3- g o o d □ 4-well□ 5- (Very well) □

5- How comfortable are you with using grammar and vocabulary accurately in your writing?

a-Very comfortable.□ b-Somewhat comfortable.□

c-Not very comfortable. □ d-Not comfortable at all. □

6- How often do you seek feedback or assistance from your teacher or peers regarding your writing?

a-Always. □ b-Often. □ c-Sometimes.□ d-Rarely.□ e-Never. □

7- How well do you think you can express your ideas and thoughts clearly in writing?

a-Very well. □ b-Moderately well. □ c-Not very well. □ d-Not well at all. □

8- How frequently do you revise and edit your writing to improve its quality?

a-Always. □ b-Often. □ c-Sometimes. □ d-Rarely. □ e-

Never. □

9- How comfortable are you with using appropriate academic or formal language in your writing?

a-Very comfortable. □ b -Somewhat comfortable. □

c-Not very comfortable. □ d-Not comfortable at all. □

10- How would you rate your overall writing competence in English compared to your other language skills (listening, speaking, reading)?

a-Stronger than other skill □ b-About the same as other skills. □

c-Weaker than other skill □

11- In your opinion, what areas of writing do you need the most improvement in?

a-grammar □ b-vocabulary□ c- organization□ d-clarity of ideas□

12- How do you typically prepare or plan your writing before you start?

a-Outlining or creating a structure. □ b-Brainstorming ideas. □

c-Researching and gathering information. □ d-Other (please specify).□

- Please provide any additional comments or insights you have about your writing competence or areas you would like to focus on for improvement:

……

……..

…

Thank you for participating in this questionnaire! Your feedback will help us understand your writing competence and improve our teaching practices.

Appendix (3) Descriptive Statistic of Teachers' Questionnaire.

minister writing assessments in your EFL classroom?											
Valid	**Missing**	**Mean**	**Standard**	**Median**	**Mode**	**Std. deviation**	**Variance**	**Range**	**Minimum**	**Minimum**	**Sum**
6	**0**	**3.333**	**3.333**	**3.500**	**4.00**	**.8165**	**.667**	**2.00**	**2.00**	**4.00**	**20.00**
2. How do you define the purpose of EFL writing assessments in your classroom?											
Valid	**Missing**	**Mean**	**Standard Error of**	**Median**	**Mode**	**Std. deviation**	**Variance**	**Range**	**Minimum**	**Minimum**	**Sum**
6	**0**	**1.833**	**.4013**	**1.500**	**1.00**	**.9831**	**.967**	**2.00**	**1.00**	**3.00**	**11.00**
3 What types of writing prompts do you typically use in EFL writing assessments?											
Valid	**Missing**	**Mean**	**Standard Error of**	**Median**	**Mode**	**Std. deviation**	**Variance**	**Range**	**Minimum**	**Maximum**	**Sum**
6	**0**	**1.666**	**.333**	**1.500**	**1.00**	**.8165**	**.667**	**2.00**	**1.00**	**3.00**	**10.00**
ou ensure that the writing prompts used in assessments are relevant and meaningful to your EFL students?											
Valid	**Missing**	**Mean**	**Standard Error of**	**Median**	**Mode**	**Std. deviation**	**Variance**	**Range**	**Minimum**	**Maximum**	**Sum**
6	**0**	**2,333**	**.333**	**2.500**	**3.00**	**.8165**	**.667**	**2.00**	**1.00**	**3.00**	**14.00**
5 How do you assess the validity of EFL writing assessments in your classroom?											
Valid	**Missing**	**Mean**	**Standard Error**	**Median**	**Mode**	**Std. deviation**	**Variance**	**Range**	**Minimum**	**Maximum**	**Sum**

6	0	2.500	.2236	2.500	2.00	.5477	.300	1.00	2.00	3.00	15.00

6 How do you address the issue of reliability in EFL writing assessments?

Valid	Missing	Mean	Standard Error of	Median	Mode	Std. deviation	Variance	Range	Minimum	Maximum	Sum
6	0	1.833	.3073	2.000	2.00	.7527	.567	2.00	1.00	3.00	11.00

7-In your opinion, what are the biggest challenges in assessing EFL writing validity?

Valid	Missing	Mean	Standar d Error	Median	Mode	Std. deviatio	Variance	Range	Minimum	Maximum	Sum
6	0	2.166	.4013	2.500	3.00	.9831	.967	2.00	1.00	3.00	13.00

8 How do you provide feedback to students based on their writing assessments?

Valid	Missing	Mean	Standar d Error	Median	Mode	Std. deviatio	Variance	Range	Minimum	Maximum	Sum
6	0	2.333	.3333	2.500	3.00	.8145	.667	2.00	1.00	3.00	14.00

9 How do you use the results of EFL writing assessments to inform your instruction?

Valid	Missing	Mean	Standar d Error	Median	Mode	Std. deviatio	Variance	Range	Minimum	Maximum	Sum
6	0	1.833	.4013	1.500	1.00	.9831	.967	2.00	1.00	3.00	11.00

10 What changes or improvements would you suggest to enhance the validity of EFL writing assessments?

Valid	Missing	Mean	Standar d Error	Median	Mode	Std. deviatio	Variance	Range	Minimum	Maximum	Sum
6	0	2.500	.3415	3.000	3.00	.8366	.700	2.00	1.00	3.00	15.00

Appendix (4) Descriptive Statistic Table of Students' Questionnaire

1. How confident do you feel about your writing skills in English?

Valid	Missing	Mean	Standard Error of	Median	Mode	Std. deviation	Variance	Range	Minimu m	Maximu m	Sum
20	0	2.000	.1622	2.000	2.000	.7254	.526	2.00	1.00	3.00	40.00

2. How often do you practice writing in English outside of the classroom?

Valid	Missing	Mean	Standard Error of	Median	Mode	Std. deviation	Variance	Range	Minimum	Maximu m	Sum
20	0	3.000	2809	3.00	4.000	1.256	1.579	4.00	1.00	5.00	60.00

. Which types of writing tasks do you find most challenging? (Select all-that apply)?

Valid	Missing	Mean	Standard Error of	Median	Mode	Std. deviatio	Variance	Range	Minimum	Maximum	Sum
20	0	1.700	.2625	1.00 0	1.00	1.174	1.379	3.00	1.00	4.00	34.00

4. On a scale of 1-5, how well do you understand the structure and organization of a well-written paragraph or essay?

Valid	Missing	Mean	Standard Error of	Median	Mode	Std. deviatio	Variance	Range	Minimum	Maximum	Sum
20	0	3.450	.2562	3.000	3.00	1.145	1.313	4.00	1.00	5.00	69.00

5. How comfortable are you with using grammar and vocabulary accurately in your writing?

Valid	Missing	Mean Error of	Standard Error of	Median	Mode deviation	Std. deviatio	Variance	Range	Minimum	Maximum	Sum

20	0	2.150	.1312	2.000	2.00	0587 1	.345	2.00	1.00	3.00	43.00

6. How often do you seek feedback or assistance from your teacher or peers regarding your writing?

Valid	Missing	Mean	Standard	Median	Mode	Std.	Variance	Range	Minimum	Maximum	Sum
20	0	3.000	.1777	3.000	3.00	.7947	.632	4.00	1.00	5.00	60.00

7. How well do you think you can express your ideas and thoughts clearly in writing?

Valid	Missing	Mean	Standard Error of	Median	Mode	Std.	Variance	Range	Minimum	Maximum	Sum
20	0	1.550	.1534	1.000	1.00	.6883	.471	2.00	1.00	3.00	31.00

8. How frequently do you revise and edit your writing to improve its quality?

Valid	Missing	Mean	Standard	Median	Mode	Std. deviation	Variance	Range	Minimum	Maximum	Sum
20	0	2.150	.2325	2.500	3.00	1.039	1.082	3.00	1.00	4.00	43.00

9. How comfortable are you with using appropriate academic or formal language in your writing?

Valid	Missing	Mean	Standard Error of	Median	Mode	Std. deviation	Variance	Range	Minimum	Maximum	Sum
20	0	2.400	.1685	2.000	2.00	.7539	.568	3.00	1.00	4.00	48.00

10. How would you rate your overall writing competence in English compared to your other language skills (listening, speaking, reading)?

Valid	Missing	Mean	Standard Error of	Median	Mode	Std. deviation	Variance	Range	Minimum	Maximum	Sum
20	0	1.900	.1235	2.000	2.00	.5525	.305	2.00	1.00	3.00	38.00

11. In your opinion, what areas of writing do you need the most improvement in?

Valid	Missing	Mean	Standar d Error	Median	Mode	Std. deviatio	Variance	Range	nimu m	ximu m	Sum
20	0	1.950	.1983	2.000	2.00	.8870	.787	3.00	1.00	4.00	39.00

12. How do you typically prepare or plan your writing before you start?

Valid	Missing	Mean	Standar	Median	Mode	Std. deviation	Variance	Range	Minimum	Maximum	Sum
20	0	2.400	.1521	2.500	3.0	.6805	078 7	2.00	1.00	3.00	48.00

d Error

Appendix (5) Writing Test Papers

Sabratha University

Faculty of Education, Zulton

English Language Department

Academic Writing / Med- Term Exam
Total Marks:40

Q1- Answer the following True(T), False(F) questions and correct the false one:

1- A successful topic sentence contains an idea about the topic.

2- A short essay has two basic parts.

3- The concluding is usually two or four sentences in length.

4- A paragraph is not a group of sentences about a single topic.

5- When arranging ideas in order of importance, use language such as in the beginning, next, then, first, second, or finally.

(5 Marks)

Q2- Complete the following sentences:

1- ……. is the last paragraph. It brings the essay to a close.

2- A……...is a ……...that begins the introduction.

3- A five paragraph essay has…… …… paragraphs.

4- …… usually comes at the end of the introduction.

5- A paragraph has ……...when all the sentences ……. the topic sentence.

(5 Marks)

Q3- Write a clause for each of these topics.

1- a favorite place to relax.

2- a grandparents

3- a pet I have Known

4- a favorite food to eat

5- playing a musical instrument

Topic 1:

……..

Topic 2:

……..

Topic 3:

……..

Topic 4:

……..

Topic 5 :

……..

(10 Marks)

Q4- Show your familiarity with the following:

1- Parts of an essay

2- Hook and background information.

3- Run-on sentences.

(10 Marks)

Q5- Write a paragraph on ONE of the following topics sentences:

1- Technology is making our lives easier.

2- A famous person whom you admire.

3- Students who both work and attend school lead busy lives.

(10 Marks)

GOOD LUCK

Sabratha University

Faculty of Education, Zulton

English Language Department

Academic Writing / Final Exam

Name……...
ID.no……...

Q1. Fill in the correct connectives and linking words from the list.

As Long as_ Because_ However_ In Spite of _ Therefore

1……. the low temperatures during the winter, Moscow is always worth visiting.

2. I can't come now ……. the children are ill, and I have to look after them.

3.The economy collapsed, ……. the government had to resign.

4.You may go out with your friends tonight, …… you never go alone and stay with them all the time.

5.We had a wonderful time in Barcelona ……, transport workers were on strike so we couldn't use the underground.

Q2. Define the following?

1.Thesis Statement

……
……
……
……
……
……

2.An Outline

……
……
……

3.Strong Introduction

1……
…….

2……
…….

3……
…….

Q3. Label each statement T for thesis statement, M for main idea, and S for supporting point.

Title: The Benefits of Yoga

a. ----------Develops clear thinking
b. ----------Physical benefits
c. ----------Improves concentration
d. ----------Reduces fear, anger and worry
e. ----------Mental benefits
f. ----------Improves blood circulation
g. ----------Improves digestion
h. ----------Helps you feel calm and peaceful
i. ----------Develops self-confidence
j. ----------Doing yoga regularly can be good for your mind, your body and your emotions.

Q4. Read these thesis statements below. Write S (if it is strong thesis statement), F (if it is fact= weak thesis statement), or N (if there is no clear opinion= very weak thesis statement)

1.School uniforms provide many benefits to students, parents and educators.
( )

2. Participating in volunteer work is essential to the development of strong character.
( )

3.I like my apartment

( )

4.Technology has changed our lives.

( )

5.Taking care of older pets can be challenging.

( )

Q5. Write a five paragraphs essay on ONE of the following topics:

1.Why do you study English? **2. The first day at school.**

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

Best of luck for all

k. Sabratha University

Faculty of Education Zulton

English Language Department

Spring 2022

Final Exam/Writing III

Hours: 2:30

Name……

……..

Q1. Complete the sentences with the connectors and transitions in the box. (10 Marks)

Also - as - even though – in addition- both – for example- for instance.

1……people prepare their model airplanes for flight, they are aware that the procedures are the same as the pre-flight procedures that pilots follow.**2**……model planes and real planes require maintenance to be operated safety.**3**……., they both need to be fueled before they can lift off. In a way, a model airplane enthusiast serves as the ground crew for the model aircraft. He or she must refuel the model plane before each flight and do a visual check of the aircraft. The model airplane flyer must **4**……...test the flight controls just as a real pilot checks the flight controls of his or her plane before take-off.

5……similarity between the two airplanes is the physics involved in flying them.**6**……., just like a real airplane, a model airplane also has wings, which create lift. This lift keeps the model floating in the air.**7**……...both planes use ailerons and flaps to control their direction. Surprisingly **8**……. A model plane is only five

feet long, it flies at about 80 mph (129 kph), which is just 20mph (32kph) slower than a real plane.

……

Q2. For each pair of sentences, check (✓) the better topic sentence. (5 Marks)

1 …… a. A person who is interviewing for a job should arrive on time to the interview.

……. b. A person who is interviewing for a job should do three important during the interview.

2 ……. a. smartphones have many useful features for communication.

……. b. smartphones are often used to send text messages.

3 ……. a. Fossils are the remains of plants or animals that died a long time ago.

……. .b. there are numerous techniques that scientists use to discover the age of a fossil.

3 …… a. there are many theories about who killed John F. Kennedy.
…… b. John F. Kennedy was assassinated on November 22, 1963.

5 ………. a. online dictionaries can help students in two important ways.
…… b. online dictionaries are available in numerous languages.

Q3. State whether the following thesis statements are weak or strong. Why? (10 Marks)

Example: Crime must be stopped.

Weak because it is a general statement. What crime? Where?

1.The court needs to implement stronger sentences.

……
……..

2.History is an important subject.

……
……...

3.Charles Dickens uses the setting of his novels to emphasize the theme of class division.

……
……

4.Socialism is the best form of government for Kenya because it will promote equal opportunity for workers.

……
……

5.If the government takes over the copper industry in Kenya, it will become more efficient through regulation and standardization.

……
……

6.Sigmund Freud is one of the greatest psychologists in medical history.

……
…….

7.Because Banana Herb Tea Supplement promotes rapid weight loss that results in the loss of muscle and lean body mass, it poses a potential danger to customers.

8.Movies are becoming more and more daring in their subject matter.

……
…….

Q4. State whether the following statements are "True " or " False"
(5 Marks)

1.A comparison essay is one of the most common forms of essay writing. In this type, the writer discusses two subjects, and these subjects can be anything, including people, objects places or ideas. ( )

2. edit your outline by removing irrelevant ideas or details.
( )

3.an essay hook is a strong opening sentence capturing readers' attention.
( )

4.the thesis statement is the sentence that tells the main idea of the whole essay.
( )

5.in the point-by-point method, the writer discusses points of comparison about one subject first before discussing the same points about the second subject.
( )

Q5. Read these ten essay titles. Put a check mark (✓) next to the most 5 appropriate for a comparison essay.
(10 Marks)

1……. Why people should be Vegetarians.

2……. Laptop Computer and Desktop Computer.

3……. Home cooking vs. Restaurant Cooking.

4…… Male Bosses and Female Bosses.

5……. The worst day of my professional or academic life.

6……. Life as an only child and life with siblings.

7……. Major personality types of young children.

8……. The steps in writing a successful resume.

9……. A comparison of the book pride and prejudice and a film version.

10……. The unforeseen effects of intercontinental travel.

Q6. Write a five – paragraphs essay on ONE of the following topics:
(20 Marks)

1. Importance of friends in our life **2. The Happiest Day in your Life.**

……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

Best of luck for all