International Journal of Language and Literary Studies
Volume 7, Issue 3, 2025
Homepage : http://ijlls.org/index.php/ijlls

Optimizing EFL Learners' Outcomes in Formative Assessment through Educational Technology Integration

Ayad Chraa*
TEFL Teacher Trainer, CRMEF Souss Massa, Morocco
prof.ayadchraa@gmail.com

Mohamed Alidrissi
EFL Teacher, Ministry of Education, Morocco
prof.mohamed.alidrissi@gmail.com

DOI: http://doi.org/ 10.36892/ijlls.v7i3.2141
Optimizing EFL Learners' Outcomes in Formative Assessment through Educational Technology Integration

APA Citation: Chraa, A., & Alidrissi, M. .. (2025). Optimizing EFL Learners' Outcomes in Formative Assessment through Educational Technology Integration. *International Journal of Language and Literary Studies. 7*(*3*).456-486. http://doi.org/10.36892/ijlls.v7i3.2141

Received:
24/04/2025

Accepted:
02/05/2025

Keywords:
Reality and Imagination
Mental spaces
Conceptual blending
Poetic iconicity
Temporal forms.

Abstract

The introduction of technology in language teaching has significantly transformed assessment practices, particularly in the domain of formative assessment. This study examines the impact of Plickers, a technology-based assessment tool, on the performance of English as a Foreign Language (EFL) learners. Situating its foundation in Kane's system of argumentative validity, the research explores how technology can enhance formative assessment by providing immediate feedback, individualized support, and data-driven insights. Utilizing an experimental design, the study involved 150 EFL learners divided into control and experimental groups. The experimental group engaged with Plickers-enhanced formative assessment, while the control group followed traditional assessment methods. Quantitative analysis revealed a statistically significant improvement in learner performance in the experimental group, with a large effect size (Cohen's d = 2.75 and 2.42). These findings demonstrate the strong potential of technology to foster more effective and dynamic formative assessment practices in EFL contexts. The study not only supports the validity of technology-integrated assessment through empirical evidence but also emphasizes its practical value in enhancing learner outcomes. The implications highlight the importance of integrating digital tools into language classrooms to meet contemporary pedagogical needs.

1. INTRODUCTION

In the field of education, formative assessment is widely recognized as a vital component of effective teaching and learning (Spector & Yuen, 2016). Particularly in the No Child Left Behind (NCLB) era, formative assessment has gained prominence as an instructional strategy that can significantly improve student achievement, especially among

low-performing learners. Popham (2011) defines formative assessment as "a planned process in which assessment-elicited evidence of students' status is used by teachers to adjust their ongoing instructional procedures or by students to adjust their current learning tactics" (p. 270). Unlike summative assessment, which occurs at the end of an instructional unit to evaluate overall learning outcomes, formative assessment is embedded within the learning process. It serves as a continuous feedback mechanism that allows educators to identify learner strengths and areas in need of development in real-time. This iterative process fosters instructional responsiveness, enabling teachers to adapt their teaching strategies and learners to refine their approaches to studying. Moreover, formative assessment supports the cultivation of a growth mindset by encouraging active learner engagement and reflection. It empowers students to take ownership of their learning progress and contributes to the creation of a more dynamic and interactive educational environment. By continuously informing both instruction and learning, formative assessment plays a pivotal role in enhancing academic performance and fostering meaningful educational development.

Stiggins (2002) asserted that “if we are finally to connect assessment to school improvement in meaningful ways, we must come to see assessment through new eyes” (p. 758). In the 21st-century educational context, the integration of technology into EFL classrooms has emerged as a critical element for enhancing both teaching quality and student learning outcomes. Despite this trend, Bhagat and Spector (2017) noted that many studies on formative assessment continue to overlook the central role of technology, indicating a persistent gap in the literature. As learners increasingly engage with digital technologies, their familiarity with and enthusiasm for digital platforms necessitate the intentional incorporation of these tools into pedagogical practices. Effective integration can harness students’ interest in technology to foster deeper engagement and more meaningful learning experiences. In line with this perspective, a growing body of research highlights the transformative role of technology in education, particularly in cultivating dynamic, learner-centered environments (Caldwell, 2007; Danielson, 2011; Irving, 2015). These studies provide substantial empirical support for the impact of technology on instructional practices, learning processes, and assessment strategies—ultimately contributing to learners' knowledge enrichment and skills development.

Among the most promising applications of educational technology is its use in formative assessment procedures. Technological tools not only facilitate real-time feedback and continuous learning assessment but also support flexible and adaptive classroom environments. As Irving (2015) observed, such tools "assist in the formative assessment process by supporting classroom environments that allow students and teachers to assess learning and providing mechanisms to present information about student learning during instructional sequences" (p. 380). This integration of technology into formative assessment represents a pivotal step toward more effective, responsive, and personalized learning experiences in EFL settings.

One of the foremost objectives in contemporary pedagogy is fostering active student engagement within the formative assessment process. To effectively gauge comprehension and address misconceptions, educators employ a range of strategies, including diagnostic assessments, unit quizzes, exit tickets, and collaborative activities such as think-pair-share. Formative assessment extends beyond the evaluation of student learning; it is also instrumental in shaping and refining instructional methods and techniques. In this way, formative assessment serves as a critical mechanism for informing pedagogical decisions and ensuring that teaching is responsive to learners' evolving needs.

The advent of accessible educational technologies has significantly expanded the possibilities for integrating formative assessment into daily classroom practices. Tools such as

Classroom Response Systems (CRSs)—including platforms like Plickers, Kahoot, and Socrative—have demonstrated considerable potential in supporting both teachers and learners. These digital tools enable the collection of real-time assessment data, allowing educators to provide immediate and targeted feedback. Such functionalities enhance the responsiveness and interactivity of formative assessment, promoting a more personalized and adaptive learning environment. As Beatty and Gerace (2009) aptly note, "Teachers have limited time to assess students' performances and provide feedback, but new advances in technology can help solve this problem" (p. 142). This insight highlights the vital role of technology in addressing practical classroom constraints while enriching the quality and efficiency of formative assessment practices.

The rationale for incorporating technological tools to enhance formative assessment in English as a Foreign Language (EFL) instruction derives from recent developments in educational technology that prioritize immediate, personalized feedback as a cornerstone of effective remedial pedagogy. From the extensive range of available platforms, Plickers was chosen due to its distinctive compatibility with resource-constrained educational settings: requiring only a single teacher-operated device while enabling student responses through printed cards, thereby ensuring both cost-effectiveness and accessibility. This characteristic renders Plickers particularly advantageous in contexts characterized by limited digital infrastructure, including numerous Moroccan public institutions. The platform facilitates rapid identification of learning deficits and provides targeted pedagogical interventions grounded in real-time learner performance data. To augment this approach, technology-enhanced remedial instruction was operationalized through digitized reinforcement activities, interactive multimedia resources, and adaptive exercises calibrated to individual learner profiles. These methodologies served not only to reinforce the formative assessment framework but also to facilitate differentiated instructional practices. This investigation seeks to advance the discourse in language assessment by elucidating the pedagogical efficacy of integrating practical, scalable digital tools within formative assessment protocols.

1.1. Literature Review

Validity Theory

Validity has long been recognized as a foundational concept in educational measurement and psychometrics, serving as the cornerstone of effective assessment design and interpretation. As Angoff (1988) notes, validity is a critical psychometric concern, essential to ensuring that assessment tools accurately measure the constructs they are intended to measure. Its centrality to the conceptual framework of testing underscores its importance in guiding the development, use, and interpretation of assessment instruments. Over time, validity theory has evolved in response to shifts in educational priorities and theoretical insights. Messick (1989) emphasized a transition from viewing validity as a set of discrete categories—commonly referred to as the "Trinitarian model" of content, criterion-related, and construct validity—to a more holistic "unitary" conception. In this integrated view, validity is no longer limited to the test itself but is fundamentally concerned with the inferences drawn from test scores and the consequences of their use. This paradigmatic shift reflects the dynamic and context-sensitive nature of validity: what is deemed valid in one context may not hold in another.

Building on this transformation, contemporary validation practices place a strong emphasis on the collection of empirical evidence to support test score interpretation and use. Urbina (2014) reinforces this view by highlighting that current psychometric theory prioritizes the meaningful application of test results over mere face-value content alignment.

Consequently, the primary concern for educators and assessment designers is not only the technical quality of the instrument but also the appropriateness of the decisions informed by it. The theoretical development of validity has undergone three major phases, reflecting its historical and conceptual evolution (Anastasi, 1986; Angoff, 1988). From early empirical validation models to the more nuanced interpretive frameworks of today, these phases illustrate the growing recognition that validity is not static, but rather an evolving construct shaped by emerging research and contextual demands. When rigorously applied, validity theory provides a powerful foundation for educational assessments that yield reliable, interpretable, and pedagogically meaningful outcomes.

Historically, the concept of validity emerged in the early 20th century as a psychometric indicator determined by the extent to which test scores accurately predict or estimate outcomes on relevant external criteria. This early approach to validity was predominantly criterion-referenced, grounded in practical measurement objectives and statistical correlations. Shaw and Crisp (2011) noted that this foundational understanding of validity was closely tied to pre-existing interest attributes, emphasizing the role of observable outcomes in shaping the epistemological and philosophical underpinnings of this psychometric era. In this context, validity was understood as context-dependent; a test was considered valid only for the specific purposes, populations, and settings for which it was designed. (Garrett & Woodworth, 1973) supported this position, arguing that the generalizability of a test is not inherent but must be cautiously inferred. This view remains relevant today, as recent scholarship has continued to emphasize the importance of context in validation processes. For instance, Haeffel and Cobb (2022) contend that test generalizability plays a crucial role in diversifying psychological research and strengthening theoretical frameworks.

Beyond contextual relevance, empirical grounding is another essential component of test validity. A valid test must produce scores that are empirically measurable and theoretically interpretable. This aligns with the operational perspective of validity, which emphasizes observable correlations with defined constructs. Guilford (1946) advanced this understanding by suggesting that a test's validity is determined by the breadth of meaningful relationships it shares with external variables. From this perspective, the extent to which a test correlates with the attributes it is intended to measure is central to its validation. Together, these views reflect the multifaceted and evolving nature of validity. What began as a criterion-based assessment of predictive accuracy has expanded into a broader, more nuanced framework that considers context, empirical evidence, and theoretical alignment.

The second major phase in the evolution of validity theory emerged during the 1950s, marked by the formalization of the content, criterion, and construct validity framework—often referred to as the trinitarian model. This period signaled a significant departure from the earlier criterion-based model, which primarily emphasized predictive accuracy, and introduced a more structured yet fragmented conceptualization of validity. The first edition of the *Standards for Educational and Psychological Testing* operationalized validity through content, predictive, concurrent, and construct dimensions. Although this classification aimed to clarify the diverse facets of test validation, it also reflected the complexity and lack of theoretical integration within the construct of validity itself. Despite the compartmentalized nature of this model, a pivotal agreement emerged across the Standards: validity is not an inherent property of a test, but rather a reflection of the appropriateness of the interpretations and uses of test scores. This shift reframed the focus of validation away from static test attributes and toward the contextual and inferential use of assessment outcomes. One of the key motivators behind this shift was the recognition of limitations within the earlier criterion-based approach. Shaw and Crisp (2011) argued that this model failed to account adequately for the relevance of test content, particularly when external criteria were insufficiently representative of the construct

being measured. To address these limitations, content validity was introduced as a means of evaluating the extent to which test items adequately sampled the domain of interest. This approach emphasized alignment between test content and instructional or conceptual domains, thereby enhancing the representativeness of assessment instruments.

Messick (1989) later reinforced this principle, identifying content-validity evidence as essential for establishing "the domain relevance and representativeness of the test instruments" (p. 17). However, this content-focused orientation was not without its shortcomings. It did not account for the internal cognitive processes employed by test takers—processes critical to understanding how individuals interact with test content. Cronbach (1971) brought attention to this issue, emphasizing that judgments about content validity must be confined to observable test characteristics. He argued that hypotheses regarding unobservable internal processes must be validated empirically through construct validation procedures, not inferred from content alone. Thus, while the trinitarian model expanded the scope of validity, it also exposed the need for a more unified and empirically grounded framework—one that considers both observable attributes and theoretical underpinnings of test performance.

The emergence of cognitive and affective criteria as essential components in the measurement of psychological and educational constructs revealed significant limitations in purely empirical approaches to validation. Many psychometricians came to recognize that traditional empirical methods were insufficient to capture the complexity of these constructs, particularly those involving internal states and processes that are not directly observable. This recognition catalyzed the development of construct validity, which sought to address the implicit criteria often overlooked in earlier models of validity.

Construct validity aimed to provide a theoretical foundation for evaluating attributes that, while not immediately observable, exert substantial influence on the accurate conceptualization and measurement of a target construct. To bridge the gap between theoretical assumptions and empirical observations, Cronbach and Meehl (1955) introduced the concept of the nomological network. This framework established a systematic structure for defining a construct in terms of its relationships with other theoretical and observable variables, thereby enhancing the coherence and depth of validation efforts.

The introduction of construct validity marked a transformative shift in how validity was conceptualized. It expanded the focus beyond the test itself to include the interpretations and uses of test scores, emphasizing that validation is an ongoing process grounded in both theoretical soundness and empirical evidence. As such, construct validity laid the groundwork for the unitary model of validity, which integrates various forms of evidence under a single, cohesive framework. This unified perspective continues to serve as the dominant paradigm in modern psychometrics, reflecting the multifaceted nature of test validation and its reliance on both conceptual clarity and methodological rigor.

Validity theory underwent a significant transformation during the 1980s and 1990s, ushering in a new phase centered on construct validity as the unifying foundation for all other forms of validity. Cronbach (1971) emphasized that construct validity is not an inherent property of the test itself but is fundamentally linked to the interpretation and intended use of test scores. He argued that the value of construct validity lies in its capacity to capture complex theoretical variables, particularly in cases where no singular external criterion exists to predict outcomes uniquely or where no clear domain of content can be sampled with certainty. As psychometricians increasingly acknowledged the importance of score interpretation in the validation process, the necessity of incorporating multiple sources of evidence became evident. This shift paved the way for a more comprehensive and integrated understanding of validity—

one that regards construct validity as the overarching framework through which all validity evidence must be interpreted.

Among the most influential voices in this transformation was Samuel Messick, whose work in the late 1980s established the foundation for the unitary concept of validity. Messick argued that construct validity subsumes all other traditional types of validity—content, criterion-related, and consequential—by integrating them into a single, coherent framework. His emphasis on the consequences of test use and the consistency of score-based inferences expanded the scope of validation to include ethical and social dimensions, reflecting a broader concern with how test results affect educational decisions and learner outcomes. Nitko (2004) later supported this perspective, reinforcing the view that construct validation must encompass a range of evidentiary sources and implications.

This reconceptualization of validity culminated in a paradigm shift that emphasized argumentation and inference as the core of the validation process. Contemporary validity theory has thus adopted an "argument-based" approach, most notably articulated by Kane (1992, 2006, 2013). Kane proposed that validation should be understood as the process of building and evaluating arguments for the interpretation and use of test scores. His framework emphasizes the coherence, plausibility, and empirical support for the proposed inferences derived from test results. This perspective was formally recognized in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999), which redefined validity as “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9). This definition not only reflects Kane’s influence but also provides researchers with a consistent, theoretically grounded framework for test validation.

The following section of this study will focus on Kane’s argument-based approach to validity, which serves as the theoretical and methodological foundation for the present investigation.

1.2. Argumentative Validity

Argument-based validity, as advanced by Kane (1992, 2006, 2013), represents a significant and contemporary development in validation theory. It foregrounds the importance of identifying key inferences and assumptions in the construction of interpretive arguments that support the use of assessment scores. Within this framework, validation is not a static judgment of a test's properties but an ongoing process of gathering and evaluating evidence to support the decisions derived from test results.There is a general consensus in the field that validation entails the systematic accumulation of evidence to evaluate the soundness of decisions made based on assessment outcomes. In educational contexts, assessments are commonly used to draw inferences about learners' competencies and knowledge, based on their performance. The argument-based approach to validation offers a structured methodology for this process, consisting of two core steps: the formulation of interpretive claims and the evaluation of the supporting evidence.

Kane (1992) explains that "The argument-based approach to validation adopts the interpretive argument as the framework for collecting and presenting validity evidence and seeks to provide convincing evidence for its inferences and assumptions, especially its most questionable assumptions" (p. 527). This approach applies broadly to both quantitative and qualitative assessments, providing a flexible yet rigorous structure for validation across diverse educational contexts. At its core, Kane’s model emphasizes prioritization, organization, and selection—elements essential to building a coherent and compelling validity argument. However, such an argument must be substantiated by a range of evidence that supports each

inferential step. In alignment with this, the Standards for Educational and Psychological Testing underscore the need for integrated evidence: "A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (AERA, APA, & NCME, 1999, p. 17). This evaluative process helps identify strengths and weaknesses in the interpretive argument, ultimately guiding improvements in assessment design and implementation. Unlike earlier models that compartmentalized different types of validity, argumentative validity synthesizes various sources of evidence to form a logical, evidence-based justification for the interpretations and uses of test scores. As such, it offers a comprehensive framework for ensuring that assessments serve their intended purposes with theoretical and empirical rigor.

Kane's argument-based approach to validation presents a robust framework for understanding how test scores can be interpreted and used meaningfully within educational and psychological contexts. At its core, this approach posits that validity is not a static characteristic of a test but rather resides in the plausibility and coherence of an interpretive argument that connects test scores to conclusions and decisions. Kane's model emphasizes that test-score interpretations must be supported by clear, appropriate, and relevant evidence, ensuring that the proposed inferences and uses are justifiable. Central to this process is the recognition that every interpretation involves a sequence of inferential steps—from observable performances to theoretical conclusions—which must be transparently articulated and critically examined. Kane (1992) argues that "*the interpretation involves an argument leading from the scores to score-based statements or decisions, and the validity of the interpretation depends on the plausibility of this interpretive argument*" (p. 527). Therefore, before collecting validity evidence, it is imperative to explicitly identify the assumptions and inferences embedded in the interpretive chain. This transparency enables researchers and practitioners to evaluate the strength of the argument and determine whether the evidence supports the intended interpretations and uses of test scores.

In this vein, the argument-based approach to validation places considerable emphasis on the systematic analysis of interpretive assumptions, particularly those that are most uncertain or potentially problematic. By doing so, it fosters a deeper and more accountable form of validation that extends beyond superficial metrics. The test-score interpretation process, according to Kane, is inherently explanatory—it seeks to ascribe meaning to test scores and articulate their implications: "*Interpretations involve meaning or explanation*" (Kane, 1992, p. 527). Each inference within this interpretive chain is grounded in underlying theoretical assumptions, often tied to construct definitions. The strength of the overall validity claim, then, rests on the degree to which these inferences are substantiated by a coherent body of empirical and theoretical evidence. Kane (1992) affirms that "*the best that can be done is to show that the interpretive argument is highly plausible, given all available evidence*" (p. 527). As such, validation becomes an evidence-based evaluation of the interpretive argument's strength, rather than a checklist of isolated indicators. Ultimately, the argument-based framework integrates all relevant types of validity evidence into a unified structure, offering a powerful tool for ensuring that assessments fulfill their intended functions with fairness, accuracy, and theoretical integrity.

According to Kane (1992), three general criteria are essential for evaluating an interpretive argument within the framework of argument-based validity: clarity, coherence, and plausibility. The first criterion, clarity, necessitates that the argument's components—including inferences, assumptions, and conclusions—are explicitly and meticulously articulated. An

argument lacking clarity may obscure the interpretive logic, leading to flawed or unsupported test-score uses. Therefore, specificity in each step of the interpretive process is paramount. The second criterion, coherence, pertains to the internal consistency and logical structure of the argument. Coherent arguments can be evaluated within a theoretical framework and, when applicable, supported by mathematically formalized models that enhance their transparency and evaluability. The third criterion centers on the plausibility of assumptions. In Kane's view, assumptions should either carry intrinsic credibility or be supported by empirical evidence, particularly when they are uncertain or under scrutiny. Weak or vague assumptions, once identified, can be strengthened through targeted research, while imprecise inferences can be reformulated with greater precision. This ongoing refinement of the argument ensures that its inferential links remain defensible and transparent.

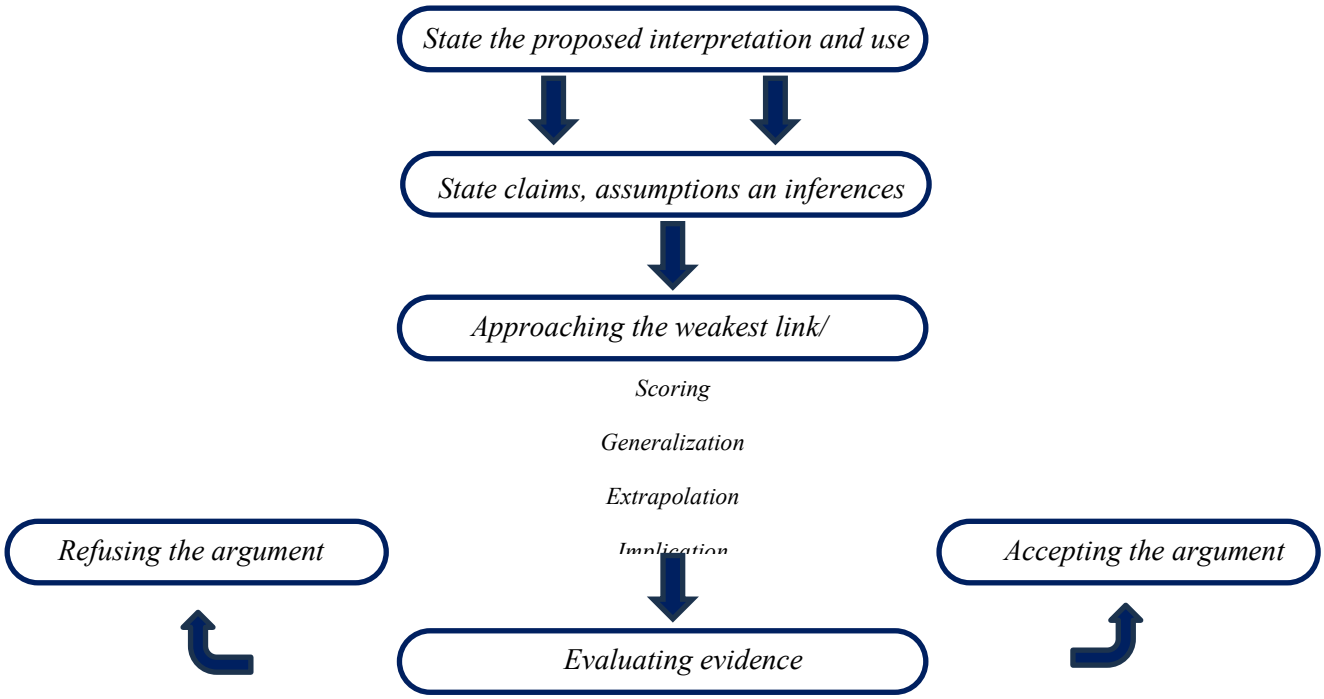
Extending this perspective, Kane underscores the importance of parallel lines of evidence and argumentation in supporting assumptions and conclusions. A conclusion corroborated through multiple lines of reasoning is inherently more resilient than one supported by a single argument. Thus, redundancy—far from being a weakness—is a strategic asset in practical validation contexts, where interpretive robustness is critical. The structure of the interpretive argument often reflects a complex web of inferences, each of which relies on underlying assumptions. Accordingly, developing multiple, independent strands of evidence to support these inferences strengthens the overall argument. Additionally, addressing and refuting plausible counterarguments plays a key role in reinforcing the validity claim. As Kane (1992) notes, "*the identification and refutation of plausible counterarguments can be a particularly effective way to reinforce practical arguments*" (p. 528). By demonstrating the implausibility of alternative interpretations, confidence in the primary argument is significantly enhanced. Therefore, the validation process should not only aim to build supportive evidence but also engage critically with potential objections. This dual focus on substantiating claims and disarming counterclaims contributes to a more resilient and persuasive interpretive argument, which lies at the heart of the argument-based approach to validation.

Kane (1992, 2006, 2013) delineates the assessment process as a sequential chain of inferences, beginning with a single observation—such as a response to a multiple-choice item, a matching task, a problem-solving question, or an entry in a portfolio. This process progresses through a series of inferential steps that culminate in real-world decision-making. The first step involves scoring, in which an observed performance is quantified with the aim of achieving accuracy, consistency, and reproducibility. From this initial scoring, the next inferential leap is generalization, wherein individual observation scores are aggregated into a test score intended to reflect broader performance. Generalization concerns the extent to which the sample of items or performances (e.g., test questions, tasks, or rater judgments) represents the larger domain of interest—the "test world." This stage critically examines whether the results obtained under specific conditions can be extended to the domain they are intended to represent. Evidence for this inference typically comes from reliability studies (Feldt & Brennan, 1989) and generalizability theory (Cronbach, Gleser, & Nanda, 1972; Brennan, 1983), which help determine the extent to which test scores are free from measurement error and robust across varying conditions of administration.

The validity of generalization relies on the assumption that minor variations in testing conditions do not significantly affect the scores—a notion captured by what Kane refers to as "invariance laws." These laws assert that the outcomes of an assessment should remain stable despite changes in factors such as item selection, rater identity, or testing environment. The strength of generalization is assessed through the design of sampling procedures and the reproducibility of results, both of which are essential for drawing defensible inferences. However, while generalization is a fundamental component of the interpretive argument, it is only one part of a broader validation chain. Following generalization, the extrapolation inference seeks to extend the test-world performance to anticipated real-world performance. Finally, the implication inference uses this extended interpretation to support specific decisions or actions based on the test results. Thus, while generalization provides a foundation for interpreting test scores, it is neither sufficient nor complete. A comprehensive validation process must incorporate all three inferential stages—generalization, extrapolation, and implication—each supported by robust empirical and theoretical evidence to uphold the integrity of score-based decisions.

Figure 1. The process of evaluating a validity argument

The extrapolation inference represents a crucial next phase in Kane's (1992, 2006, 2013) argument-based validation framework, wherein test performance is connected to real-world performance. While generalisation links observed performances to a broader universe of test-world tasks, extrapolation extends this connection further, bridging the gap between performance in the controlled test environment and anticipated outcomes in authentic, real-life contexts. This transition is essential, as it moves beyond the internal structure of the test to



address whether the scores meaningfully reflect the examinee's ability to perform in target real-world situations.

Support for the extrapolation inference is typically drawn from two main sources of evidence. First, one must demonstrate that the test content meaningfully represents key dimensions of real-world performance. This may involve designing tasks that authentically simulate real-life scenarios. Second, empirical analyses—such as correlation studies—are used to establish relationships between test scores and performance in practical, non-test settings.

The strongest form of extrapolation evidence emerges when test scores show high correspondence with external assessments that are theoretically aligned with the construct being measured.

However, generalisation and extrapolation can be in tension with one another. Kane (2006) explains that "we can strengthen extrapolation at the expense of generalisation by making the assessment tasks as representative of the target domain as possible, or we can strengthen generalisation at the expense of extrapolation by employing larger numbers of highly standardised tasks" (p.37). This trade-off illustrates the complexity of assessment design: highly standardised formats may enhance reliability and generalizability, yet may fall short in replicating real-world performance. Conversely, more authentic tasks may better support extrapolation but introduce variability that undermines generalisation. Thus, achieving a defensible balance between these inferences is vital to ensuring the overall validity of test score interpretations.

The final stage in Kane's argument-based approach to validation involves the implication inference, which proceeds from a test score to its interpretation, and subsequently, from that interpretation to a decision or course of action. This stage emphasizes that valid score interpretation alone is insufficient; what ultimately matters is whether the decisions based on those interpretations lead to beneficial, ethical, and justifiable outcomes. In this regard, the accurate measurement of a construct does not guarantee the practical usefulness of the information it yields. Kane (2006) stresses this point, asserting that "a decision procedure that does not achieve its goals, or does so at too high a cost, is likely to be abandoned even if it is based on perfectly accurate information" (p. 61). Thus, validity must also consider the consequences of test use, particularly how well decisions serve their intended purposes and align with societal and educational values.

Building upon the foundational work of Guion and Messick, Kane argues that the legitimacy of test use depends on more than just technical adequacy; it requires critical examination of the assumptions about the intended outcomes and the value systems embedded in those outcomes. In other words, evidence supporting the accuracy of score interpretation does not automatically warrant its application. As Kane (2006) cautions, "it is generally inappropriate to assume that evidence supporting a particular interpretation of test scores automatically justifies a proposed use of the scores" (p. 61). Therefore, the final component of the validity argument entails evaluating the broader impact of assessment practices—not only on individual learners, but also on educators, institutions, and society as a whole. This impact-focused perspective ensures that assessment remains aligned with ethical standards and contributes positively to educational development and decision-making processes.

In summary, validation is best understood as a comprehensive process rather than a static outcome. A validated test does not merely represent a tool with accurate results; rather, it signifies the successful implementation of a systematic validation process that includes the interpretation, intended use, and context of the assessment. Validation begins with an explicit and well-defined statement of the proposed interpretation and use of test scores. This is followed by the construction of an interpretive argument, a conceptual framework that encompasses the logical sequence of inferences and the assumptions upon which they rest. The process continues through the collection, analysis, and integration of empirical evidence, all of which support the coherence and credibility of the resulting validity argument. In this framework, educators and test developers are urged to concentrate particularly on the weakest assumptions, recognizing that the strength of the overall validity argument is limited by its most vulnerable inferential link. While evidence supporting scoring, generalization, and extrapolation is often relatively strong—grounded in reliability and empirical correlations—

greater scrutiny is typically required when moving toward implications and decisions, which are the final and most consequential inferences within the argument-based validation model. These ultimate inferences carry the weight of educational decision-making and policy implications, and thus demand careful justification to ensure ethical, effective, and contextually appropriate use of test results.

1.3. Formative assessment

In response to the accountability demands introduced by the No Child Left Behind (NCLB) Act, a wide range of educational interventions have been developed to improve student achievement, with formative assessment emerging as a key strategy. Formative assessment is widely recognized for its potential to enhance both teaching quality and student outcomes, particularly for low-performing learners who benefit from targeted instructional interventions. Black and Wiliam (2018) define formative assessment as "the process by which teachers use assessment evidence to inform their teaching" (p. 3). In an earlier foundational work, they also characterized it more broadly as "all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged" (Black & Wiliam, 1998, p. 10). In alignment with this perspective, the Formative Assessment for Students and Teachers (FAST) initiative describes formative assessment as an embedded process within instruction, designed to elicit timely feedback that can be used to adjust teaching and learning in real-time to align with specific instructional goals.

Popham (2006) reinforces this definition by emphasizing that an assessment qualifies as formative only when the information it provides is used during the same instructional cycle to modify pedagogical strategies and directly support students' learning needs. He later refined this conceptualization, describing formative assessment as a deliberate, evidence-based process by which both teachers and students make informed decisions to enhance ongoing instruction (Popham, 2008). The diversity in definitions of formative assessment reflects its flexibility and range of applications. It can serve as a reflective tool for teachers and learners, or as a data source for institutional decision-making and policy development. Black and Wiliam (1998) note that formative assessment can serve a wide spectrum of feedback-related functions, including diagnosing learning difficulties, predicting future performance, and evaluating the effectiveness of instructional strategies. This broad applicability underscores the central role formative assessment plays in contemporary educational reform and performance-based instruction.

Brookhart (2007) defines formative assessment as "formative classroom assessment gives teachers information for instructional decisions and gives pupils information for improvement" (p. 43). This definition emphasizes the dual function of formative assessment: it informs both the teacher's instructional strategies and the student's learning process. Brookhart identifies three essential elements as key drivers of effective formative assessment. First, it serves as a reflective tool that informs current pedagogical practices. Second, it provides a foundational basis for making informed instructional decisions. Third, it offers students scaffolded support to understand how to improve their performance, encouraging both academic development and learner autonomy. These interconnected elements position formative assessment as a dynamic mechanism within the teaching-learning cycle, enabling real-time feedback and adaptive instruction.

Building on this perspective, López and Sicilia (2017) conceptualize formative assessment as a process through which educators provide learners with ongoing feedback during instruction. This feedback is not merely evaluative but formative in its function—it

supports learners in making the necessary adjustments to refine their understanding, advance their learning goals, and foster self-regulation. Thus, formative assessment is portrayed as both a strategic instructional practice and a learner-centered approach that cultivates reflective engagement and continuous improvement. The alignment between Brookhart's and López and Sicilia's perspectives underscores the evolving consensus that formative assessment is integral not only to effective teaching but also to the development of independent, self-directed learners.

Assessment lies at the core of the teaching and learning process—not only as a tool for measuring student progress but also as a means of providing instructional support. The conceptual distinction between formative and summative approaches has evolved over decades, beginning with Michael Scriven's (1967) introduction of the term formative evaluation in the context of teaching and curriculum development. Scriven described formative evaluation as the ongoing appraisal of educational programs, emphasizing continuous improvement rather than terminal judgment. Shortly thereafter, Bloom (1969) expanded the notion by reframing formative evaluation within the context of classroom assessment. By 1971, Bloom had popularized the term formative assessment in its modern sense—an approach focused on informing and adjusting instruction based on student learning evidence. In contrast, summative evaluation was defined as the evaluation of instructional units and curricula with the purpose of certifying achievement or measuring program effectiveness. However, the persistent objective of enhancing teaching practices and improving student learning experiences underscored the importance of formative assessment, which required the active involvement of students, teachers, and curriculum designers.

From this theoretical foundation, modern scholars have emphasized the distinct pedagogical value of formative assessment. López-Pastor and Sicilia (2017) argue that formative assessment provides immediate, actionable feedback that cannot be achieved through summative methods alone. Similarly, Robertson (2019) supports the view that formative practices significantly enhance student learning and contribute meaningfully to summative performance. Interestingly, this relationship is not unidirectional; summative assessment itself can be harnessed to inform formative goals, as suggested by Bell and Cowie (2000), who advocate for an integrated perspective on assessment. Nevertheless, the development of formative assessment has not relied solely on new tools; it also demands a transformation in classroom practices to ensure the effective use of those tools. The shift from tool adoption to pedagogical alignment between formative and summative assessment is essential in mitigating the dominance of summative pressures and enabling educators to fully realize the potential of assessment to improve teaching and learning outcomes.

Black and Wiliam (1998) provide one of the most accessible and widely cited explanations of formative assessment, framing it through the dual lens of "what" and "when." The "what" refers to the assessment-related activities carried out by both teachers and learners to evaluate and reflect on learning progress. These activities generate actionable data, which becomes valuable feedback for modifying teaching and learning strategies. The "when" underscores the importance of timing—assessment becomes formative when the information gathered is used in real-time to adjust instruction to better meet students' needs. In this regard, Wiliam and Leahy (2007) emphasize that the essence of formative assessment lies in continuous readjustment. It is not a one-off intervention, but rather a daily pedagogical routine that informs instructional decisions, as echoed by Andrade and Heritage (2017). This dynamic, responsive process aligns with Bloom's (1969) assertion that the primary purpose of formative assessment is to provide timely feedback and corrective measures at various stages of the learning process, in contrast to summative assessment, which aims to judge learner performance at the end of a unit or course.

This distinction becomes even more meaningful when considering how assessment results are practically applied. As noted by Broadfoot, Weeden, & Winter (2002) and Wiliam and Thompson (2008), any assessment tool can serve a formative purpose if its results are used to adjust instruction during the learning process—regardless of the assessment's original intent. This perspective broadens the applicability of formative assessment, highlighting its functional flexibility. Feedback, in this context, operates as a reinforcement-based mechanism—a crucial bridge that connects instructional stimuli to student responses (Haughney et al., 2020). According to Mäkipää and Hildén (2021), the intentional and strategic use of feedback is not just beneficial but essential for enhancing student performance. Thus, feedback serves as both a reflective and a motivational tool, anchoring the formative assessment cycle and promoting learner growth through iterative improvement.

Achieving meaningful progress in classroom practice is a complex and demanding endeavor, requiring sustained commitment and pedagogical transformation. Black and Wiliam (2009) emphasized that many teachers struggle to implement formative assessment effectively, particularly when it comes to fostering classroom dialogue. They argue that this challenge stems from the fact that successful formative assessment often demands a radical shift in teaching style, disrupting traditional instructional routines. Drawing from their extensive work with educators in both the United Kingdom and the United States, Black and Wiliam reflected on the practical realities of promoting formative assessment in schools. Their findings revealed that when learners are actively engaged in dialogue and peer group discussions, the classroom becomes a space for collaborative social learning, where students participate meaningfully in constructing knowledge (Black & Wiliam, 2009). Such practices, however, require intentional design and a rethinking of both instructional goals and teacher-student dynamics.

In light of these findings, educational policymakers and stakeholders are urged to invest in the professional development of teachers to build classroom environments that prioritize thoughtful questioning, reflective discourse, and learner agency. Establishing such a culture is not merely a pedagogical recommendation but a necessary shift aligned with the complexities of contemporary, postmodern educational contexts. A nuanced understanding of the challenges posed by modern society should serve as a catalyst for rigorous, evidence-based initiatives led by policymakers, practitioners, and school administrators. These initiatives must be grounded in research that explores the transformative potential of formative assessment, not only as a tool for improving academic performance but also as a key driver of lifelong learning and critical engagement.

1.4. Technology in Formative Assessment

Integrating technology into formative assessment practices has gained increasing importance, particularly as educators face challenges in implementing effective feedback mechanisms. Among the most prominent obstacles is the issue of delayed feedback, which can significantly diminish the instructional value of formative assessment. In this context, technology emerges as a powerful tool that enhances immediacy, efficiency, and engagement in the feedback process. Digital platforms enable English as a Foreign Language (EFL) teachers to deliver timely, personalized feedback, thereby minimizing the instructional burden often associated with traditional methods. The immediacy enabled by technology not only facilitates self-verification for learners but also helps align instructional content with individual learning needs and reinforces goal-oriented thinking.

Recent research has explored the role of technology in improving formative assessment practices and student outcomes. For example, Elmahdi, Al-Hattami, and Fawzi (2018) underscore the importance of centering learner engagement within the design of formative

assessments while also preparing students for summative evaluations. Their findings reveal that digital tools not only save time and instructional effort but also significantly alter student participation, allowing for more responsive learning environments. Additionally, these tools reduce the feedback turnaround time for both students and instructors, thereby enhancing the overall effectiveness of the assessment process. Such evidence affirms the transformative potential of educational technology in reshaping formative assessment to be more adaptive, learner-centered, and impactful in EFL contexts.

Robertson et al. (2019) emphasize the multiple benefits that technology brings to formative assessment practices, particularly in fostering more effective and engaging learning environments. One of the primary advantages identified is technology's ability to capture student attention and create a motivational learning climate. Additionally, they underscore the role of technology in facilitating immediate feedback, highlighting immediacy as a central component of effective formative assessment. Timely feedback not only reinforces student understanding but also enables instructional adjustments that align with learners' evolving needs. In this vein, Bhagat and Spector (2017) assert that technology delivers constructive, real-time feedback that supports the learning process, noting that any delays in feedback can negatively influence the pace and quality of learning.

Beyond feedback delivery, technology also plays a vital role in the collection and analysis of formative assessment data. These functionalities allow educators to monitor student progress more precisely and make data-informed instructional decisions. Supporting this view, Dakka (2015) affirms that digital tools enable efficient data gathering and interpretation, equipping teachers with meaningful insights into both individual and group performance. This capacity for real-time diagnostic assessment positions technology not merely as a supportive aid, but as an integral component in optimizing the effectiveness and responsiveness of formative assessment.

The integration of technology into formative assessment practices has been widely recognized for its potential to enhance instructional efficiency and student engagement. Beatty and Gerace (2009) underscore the time constraints teachers face in assessing student performance and delivering timely feedback, noting that "teachers have limited time to assess students' performances and provide feedback, but new advances in technology can help solve this problem" (p. 142). Similarly, Irving (2015) emphasizes the role of technology in creating classroom environments that facilitate ongoing assessment, highlighting its utility in offering platforms that allow both teachers and students to monitor and reflect on learning throughout instructional sequences (p. 380).

Several empirical studies further reinforce the value of digital tools in formative assessment contexts. For instance, Habler et al. (2016) assert that the incorporation of tablets in classroom settings introduces flexibility and ease, enriching the learning experience by streamlining instructional processes. Building on this, Dalby and Swan (2018) investigated the use of iPads by six mathematics teachers and found that these tools significantly enhanced formative assessment practices by offering diverse strategies that support real-time feedback and student-centered learning. Likewise, Chazan, Olsher, & Yerushalmy (2016) explored how technology supports monitoring student learning trajectories, particularly in mathematics education. Their findings revealed that digital tools facilitate the creation of interactive learning environments and foster continuous improvement by allowing teachers to adapt instruction based on immediate learner data. Collectively, these studies highlight the transformative potential of technology to reinforce formative assessment and promote more effective, personalized, and responsive teaching and learning experiences.

The International Association for the Evaluation of Educational Achievement (IEA) has increasingly embraced computer-based assessment (CBA) in its global educational studies, reflecting a broader shift toward the integration of technology in assessment practices. This transformation stems from the growing recognition that an engaging, adaptive learning environment is essential for enhancing educational outcomes. The widespread adoption of CBA is driven by its multiple pedagogical and operational advantages, including real-time data capture, individualized feedback, and increased efficiency in administration and scoring. As Herold (2016) noted, "digital devices, software, and learning platforms offer a once-unimaginable array of options for tailoring education to each individual student's academic strengths and weaknesses, interests and motivations, personal preferences, and optimal learning pace.". Jamieson and Musumeci (2017) similarly argue that technology's core function in education is to facilitate teaching and learning, making it a central pillar in the evolution of contemporary assessment frameworks. However, this growing reliance on technology has also led to concerns about the diminishing relevance of traditional paper-based assessments, which some fear may become obsolete in technologically advanced classrooms.

Despite these innovations, significant disparities persist across nations in terms of technological infrastructure and digital readiness. Many countries still lack the resources necessary to fully implement CBA, thereby maintaining conventional assessment as a foundational component of their educational systems. Nonetheless, even in traditional contexts, technology is increasingly integrated into assessment design, particularly in mitigating common limitations of paper-based tests. For instance, in constructed response items, digital tools can eliminate errors caused by illegible handwriting, which often lead to inaccurate or biased scoring. The ability to ensure legibility enhances scoring accuracy and fairness, while also broadening the range of constructs that can be assessed. Certain competencies—such as interactive problem-solving or real-time data manipulation—are inherently more accessible through digital assessment platforms than through traditional formats. As a result, technology not only improves logistical aspects of assessment but also expands the scope of what can be validly and reliably evaluated, marking a significant advancement in modern educational measurement.

2. METHODOLOGY

2.1. Research Design

The present study adopts a true experimental research design to rigorously investigate the proposed hypothesis and determine causality through statistical analysis. This methodological approach is recognized for its precision and scientific reliability, as it relies on empirical evidence to confirm or refute the initial assumptions. Among various experimental designs, true experimental design is considered the most robust due to its ability to establish clear cause-and-effect relationships. As Creswell (2019) asserts, true experiments are characterized by their methodological rigor, particularly in their ability to accommodate varying conditions while maintaining internal validity.

A defining feature of this design is the random assignment of participants to control and experimental groups, which precedes the application of the intervention. This process ensures that observed effects can be attributed to the treatment rather than to extraneous variables, thereby eliminating alternative explanations (Creswell, 2019). Moreover, true experimental designs involve the manipulation of independent variables, systematic control of conditions, and the observation of changes in the dependent variable, allowing researchers to draw strong inferences regarding causality. According to Leedy and Ormrod (2010), randomization plays a crucial role in reducing threats to internal validity by neutralizing the influence of chance differences between groups. Thus, the true experimental framework

employed in this study provides a sound foundation for evaluating the impact of technological intervention in formative assessment with a high degree of scientific rigor.

2.2. Context of the Study

Achieving a deeper understanding of the subject matter and adapting instruction to meet learners' diverse needs are among the ultimate goals pursued by educators. In this context, formative assessment has emerged as a critical tool, primarily due to the dynamic feedback loops it creates between teachers and students (Heritage, 2010). These continuous feedback mechanisms empower instructors to modify their teaching practices in real time, fostering more personalized and effective learning experiences. In parallel, the integration of educational technology in today's digital age is no longer optional—it is essential. As Bonk and Graham (2006) note, digital tools significantly enhance learner engagement and provide globally accessible, flexible learning opportunities.

Despite the increasing incorporation of technology in education, there remains a notable gap in structured frameworks for evaluating its effectiveness in instructional contexts. Means, Toyama, Murphy, Bakia, and Jones (2010) emphasize the need for more experimental and longitudinal research designs to better assess the impact of digital tools on teaching and learning. In response to this research gap, the current study employs a true experimental design to examine the effectiveness of integrating educational technology in improving students' formative assessment performance. Data for this study were collected from Ennour High School, a secondary institution located in the Souss-Massa region of Morocco. The study seeks to offer empirical insights into how digital tools can be leveraged to enhance assessment practices and support learner achievement in technology-enhanced educational environments.

2.3. Participants

In terms of demographic distribution, the participants in this experimental study were drawn from two distinct academic streams through a process of random class selection, resulting in the implementation of two separate experiments. A total of 101 Common Core students (first-year high school) from six classes participated in the study. These students were selected from a Moroccan secondary educational institution and were evenly distributed across the Arts and Science streams to ensure representative sampling.

The first experiment included 47 students from the Arts stream, accounting for 46.53% of the total participant pool. The second experiment comprised 54 students from the Science stream, representing 53.47% of the total sample. The selection of participants from multiple classes and streams was designed to capture a diverse cross-section of learners and enhance the study's internal validity. This demographic configuration provided a solid foundation for examining the effects of the experimental intervention across varied academic contexts within the same educational level.

2.4. Sampling Procedure

This study adopts a structured convenience sampling method to enhance efficiency and provide sufficient statistical power to yield significant and reliable conclusions. The selection of educational institutions and participant classes was based primarily on their availability and accessibility, which aligns with the defining principles of convenience sampling. As Rahi (2017) explains, convenience sampling involves the collection of data from a population that is readily reachable and accessible to the researcher, making it particularly suitable when logistical or resource-related constraints are present.

While it is acknowledged that non-probability sampling methods such as convenience sampling may limit the generalizability of findings, they remain effective when targeting populations that are difficult to access due to practical limitations. Given the time and effort

required to engage multiple high schools across different regions, the use of convenience sampling was deemed both practical and appropriate for the scope of this study. Despite its limitations, this method allowed for the efficient inclusion of relevant participants and facilitated the structured implementation of the experimental procedures.

2.5. Instruments

Description and Rationale

In experimental research, pre-tests and post-tests serve as essential instruments for measuring changes in the dependent variable and, ultimately, for evaluating the validity of the hypothesis under investigation. Their implementation provides a clear baseline and an outcome reference point, enabling researchers to determine the effectiveness of the intervention with greater accuracy. As such, pre- and post-testing constitutes a foundational component of rigorous experimental design, ensuring that any observable changes in outcomes can be attributed directly to the experimental conditions rather than to extraneous variables.

Moreover, the use of testing procedures reinforces the credibility, accuracy, and reproducibility of the research findings. Within this study, the tests were strategically applied across experimental and control groups to allow for valid comparisons. The control groups, which did not receive the technological intervention, served as critical benchmarks to evaluate the impact of the treatment on the experimental groups. This comparative design aligns with established research protocols emphasizing the importance of replication and internal validity in experimental studies. As echoed in the broader research community, reproducibility through repeated and controlled testing enhances the trustworthiness of results and strengthens the overall contribution of the findings to the educational research field.

Validity

The development of the pre-test and post-test instruments in this experimental study adhered to a rigorous and structured process, as the integrity of the research design directly influences the validity and reliability of the findings. By ensuring methodological rigor, the study minimizes the risk of errors and confounding variables that could compromise internal validity. In line with the standards of quantitative research, which prioritize replicability and control, the test instruments comprised predefined questions and fixed-response formats, consistent with recommendations by Nunnally and Bernstein (1994) and Foxcroft and Roodt (2013).

Since quizzes are widely recognized as effective tools of formative assessment, the structure and content of both the pre-test and post-test were developed in strict accordance with the official Moroccan ministerial guidelines for assessment. The test focused on a single construct—the language component—which was subdivided into three key subcomponents: vocabulary, grammar, and functions. To ensure content validity, the test covered three instructional units that had been delivered uniformly to all participating students as part of the same curriculum prior to the experiment. Furthermore, a pilot test was conducted with 25 first-year high school (Common Core) students from the science stream at Ennour High School to assess the clarity, coherence, and difficulty level of the items. A reflective section containing three feedback questions was appended to the quiz, inviting students to evaluate the clarity, feasibility, and difficulty of the exercises. Results from the pilot indicated that 16% of students (n = 4) found the overall quiz challenging, while 40% (n = 12) reported difficulty with the first vocabulary task. Consequently, revisions were made to the wording and structure of this task, including the addition of contextual cues and prompts to enhance comprehensibility and support student engagement.

Reliability

Ensuring the reliability of both the findings and the data collection instruments is a fundamental aspect of establishing the credibility and scientific rigor of experimental research. Reliability testing is essential for confirming the stability, consistency, and reproducibility of results, thereby reinforcing the trustworthiness of the study's conclusions. While traditional measures of rater agreement can provide basic insights, they often fall short in precisely quantifying interrater reliability. In contrast, Cohen's Kappa (κ) is widely regarded as a more accurate and robust metric for evaluating agreement beyond chance (McHugh, 2012).

Accordingly, Cohen's Kappa was computed using SPSS to assess the reliability of the data, the consistency of the data collection instrument, and the level of agreement between independent raters responsible for coding student responses. This approach ensured that the study's findings were grounded in systematically evaluated data, free from random errors or subjective interpretation. To secure the objectivity and reliability of the coding process, a category schema was developed based on clearly defined criteria derived from the Moroccan Ministerial Circular No. 175, and aligned with the structure of the assessment quiz. In the first experiment, the κ value was 0.90 for the pre-test and 1.00 for the post-test. Similarly, in the second experiment, the κ value was 0.91 in the pre-test and 1.00 in the post-test. These values reflect "almost perfect agreement" (Landis & Koch, 1977), thereby confirming the high interrater reliability and further enhancing the validity and credibility of the research outcomes.

2.6. Procedure

This study was structured into three critical phases across the implementation of its two experiments. The first phase involved the design and piloting of the pre-tests, ensuring content validity and structural clarity. These pre-tests were meticulously developed based on official curricular guidelines and then piloted to refine their rigor and effectiveness. The study was conducted at Ennour High School and comprised two experimental groups and two control groups, each receiving parallel versions of the test to maintain comparability. The control groups completed the assessments using the traditional paper-and-pencil format, without any technological intervention. In contrast, the experimental groups completed the same tests using Plickers cards, a classroom response technology. The original test content—five short exercises totaling 40 questions—was digitized and administered through Plickers to evaluate the impact of the technological format on formative assessment performance.

The second phase of the study focused on remedial instruction, which followed the analysis of pre-test data. This phase aimed to address specific areas of weakness identified in student performance, in line with the core objective of formative assessment. While both groups received remedial instruction, the experimental groups received it via technology-enhanced methods, whereas the control groups followed conventional instructional strategies.

Finally, the third phase involved administering the post-tests, which mirrored the design, content, and procedural conditions of the pre-tests. The purpose of the post-tests was to measure the effectiveness of the intervention, thereby determining whether the use of technology in formative assessment had a statistically significant impact on learners' performance.

2.7.Data Analysis

To evaluate the effectiveness of the technological intervention on EFL learners' formative assessment performance and to address the core research question, the collected data were analyzed using SPSS Statistics software. Two primary statistical tests were employed to examine the impact of the intervention.

First, the Paired Sample t-test was applied to compare the pre-test and post-test scores within each group—both experimental and control. This test aimed to determine whether the observed changes in performance within each group were statistically significant, thus assessing the internal impact of the intervention over time.

Second, the Independent Samples t-test was used to compare the post-test scores between the experimental and control groups. This analysis allowed for the identification of any statistically significant differences in outcomes between learners who experienced the technology-enhanced formative assessment and those who followed conventional methods.

Together, these tests provided a robust statistical framework for hypothesis testing by distinguishing genuine performance improvements from changes that might be attributed to random variation. The results from both tests contributed significantly to validating the effectiveness of the technological tools employed in the experimental design.

3. RESULTS

As outlined in the preceding section, the experimental design of this study necessitated the application of two statistical procedures: the Paired Sample t-test in the initial phase, and the Independent Samples t-test in the subsequent phase. These analyses were employed to accurately address the central research question: *To what extent does technology integration optimize EFL learners' performance in formative assessment?*

Prior to conducting the Paired Sample t-test, it was essential to assess the assumption of normality, which stipulates that the differences between paired observations should follow a normal distribution. This assumption is critical to ensure the validity and reliability of the t-test's inferential outcomes. The normality of the data was examined using the Shapiro-Wilk test, and the resulting p-values (i.e., Sig.) were all greater than the commonly accepted alpha level of .05 for both the control and experimental groups (see Table 1). As a result, the null hypothesis of normality could not be rejected, indicating that the assumption was satisfied and the data were normally distributed. Given this, the conditions were met to proceed with the Paired Sample t-test, the results of which are presented in Tables 2 and 3.

Table 1.

Results of the Shapiro-Wilk Test of Normality

		Statistic	df	Sig.
1st exp	CG	.952	24	.292
	EG	.919	23	.064
2nd exp	CG	.935	26	.101
	EG	.943	28	.134

Table 2.

Results of the Paired Sample T-test for pre-post-tests within group in the 1st experiment

	N	*M*	*SD*	*SE*	*MD*	*t*	*df*	*p*	*Effect Size*

CG	pre-post	24	10.20	2.9961	.6116	-1.8125	-4.383	23	0.000	Cohen's d .895
			12.02	3.4151	.6971					
EG	pre-post	23	16.17	2.823	.589	-2.957	-7.484	22	0.000	Cohen's d 1.561
			19.13	1.217	.254					

To assess within-group improvements, a Paired Samples t-test was conducted for both the control group (CG) and the experimental group (EG) in the first experiment at Ennour High School. For the control group, participants demonstrated a moderate improvement from the pre-test (M = 10.20, SD = 2.996) to the post-test (M = 12.02, SD = 3.415). The analysis yielded a statistically significant difference between the two time points, $t(23) = -4.383$, $p < .001$, indicating that the observed increase in post-test scores was not due to random chance. The effect size, measured using Cohen's *d*, was 0.89, suggesting a strong effect of the intervention in the control condition.

In contrast, the experimental group—who received the technological intervention—showed a more pronounced improvement, moving from a pre-test mean of 16.17 (SD = 2.823) to a post-test mean of 19.13 (SD = 1.217). The resulting paired samples t-test was also statistically significant, $t(22) = -7.484$, $p < .001$, with an effect size of $d = 1.56$, reflecting a very strong effect. These findings underscore a substantial gain in performance due to the intervention. Therefore, the directional hypothesis is supported, indicating that the experimental group, which utilized technology-integrated formative assessment, exhibited significantly greater improvement than the control group.

Table 3.

Results of the Paired Sample T-test for pre-post-tests within group in the 2nd experiment

		N	***M***	***SD***	***SE***	***MD***	***t***	***df***	***p***	***Effect Size***
CG	pre-post	26	10.76	3.6173	.7094	-2.0192	-5.560	25	0.000	Cohen's d 1.090
			12.78	3.9094	.7667					
EG	pre-post	28	16.57	2.673	.505	-2.964	-8.193	27	0.000	Cohen's d 1.548
			19.53	.922	.174					

The same statistical procedure was applied to the data from the second experiment to examine within-group differences in performance for both the control and experimental groups. For the control group, the pre-test mean score was 10.76 (SD = 3.617), and the post-test mean increased to 12.78 (SD = 3.909), with a mean difference of -2.02. The paired samples *t*-test revealed a statistically significant improvement, $t(25) = -5.560$, $p < .001$, confirming that the increase in post-test scores was meaningful. The effect size, calculated using Cohen's *d*, was 1.09, indicating a strong effect of the conventional intervention.

For the experimental group, who received the technology-based formative assessment treatment, a pre-test mean of 16.57 (SD = 2.673) and a post-test mean of 19.53 (SD = 0.922) were observed. The resulting mean difference of -2.96 highlights a substantial improvement in post-test scores. The paired samples *t*-test indicated a highly significant difference, $t(27) = -8.193$, $p < .001$. The corresponding effect size was $d = 1.54$, which represents a powerful effect. These findings align with the results from the first experiment and further support the rejection of the null hypothesis. Thus, the technology-integrated intervention consistently demonstrated a substantial impact on learners' formative assessment performance.

Overall, both the control groups (CGs) and experimental groups (EGs) demonstrated statistically significant improvement, as evidenced by the low *p*-values obtained through the paired samples *t*-tests. However, the experimental groups—who were exposed to the technology-integrated formative assessment—showed notably higher mean differences and larger effect sizes compared to their control counterparts. This outcome suggests that while formative assessment in any form contributes to learning gains, the integration of educational technology significantly enhances its effectiveness. The statistical findings highlight the potential of technological tools to optimize learners' performance more effectively than conventional methods.

Having examined the progression from pre-test to post-test within each group, the next analytical phase involves applying the independent samples *t*-test. This test evaluates whether the post-test performance differs significantly between the control and experimental groups within each experiment, thereby providing further evidence on the impact of technology-enhanced formative assessment (see Table 5).

Table 4.

Results of Levene's Test of Homogeneity

	F	***Sig.***
1st Exp	10.133	.003
2nd Exp	24.488	.000

Before interpreting the results of the independent samples *t*-test, Levene's Test for Equality of Variances was conducted to assess the assumption of homogeneity of variances between the control and experimental groups in each experiment. This test is evaluated against the conventional alpha level of 0.05. The results (see Table 4) indicate F-statistics of $F(1, 45) = 10.133$ and $F(1, 52) = 24.488$, with corresponding *p*-values of 0.003 and 0.000, respectively. As both *p*-values are below the alpha threshold ($p < .05$), the assumption of equal variances is violated, allowing for the rejection of the null hypothesis of homogeneity.

Given this violation, the analysis proceeds using Welch's *t*-test, a robust alternative that does not assume equal variances between groups. This methodological shift ensures the accuracy and appropriateness of the statistical comparison between the control and experimental groups' post-test scores (see Table 5).

Table 5.

Results of the control group and experimental group post-tests differences in each experiment

	N	***M***	***SD***	***SE***	***df***	***t***	***p***	***Mean Difference***	***Effect Size***

Exp 1	CG	24	12.02	3.415	.697	28.972	-9.583	0.00	-7.1096	Cohen's d 2.750
	EG	23	19.13	1.217	.254					
Exp 2	CG	26	12.78	3.909	.766	27.582	-8.581	0.00	-6.7473	Cohen's d 2.418
	EG	28	19.53	.922	.174					

An independent samples *t*-test was conducted to examine the effect of technological intervention on learners' performance across two distinct experiments. Each experiment compared a control group that underwent conventional formative assessment with an experimental group exposed to a technology-integrated approach, thereby assessing the impact of technology-enhanced formative assessment on optimizing student learning outcomes.

In the first experiment, a statistically significant difference was observed between the control group (M = 12.02, SD = 3.42) and the experimental group (M = 19.13, SD = 1.22), $t(28.97) = –9.583$, $p < .001$. The effect size, calculated using Cohen's d, was $d = 2.75$, which indicates an extremely strong effect. Similarly, in the second experiment, the post-test scores showed a significant difference between the control group (M = 12.78, SD = 3.91) and the experimental group (M = 19.54, SD = 0.92), $t(27.58) = –8.581$, $p < .001$. The effect size was also substantial, with d = 2.42, suggesting a very strong impact of the technological intervention. These findings provide compelling evidence for the efficacy of technology-integrated formative assessment in enhancing EFL learners' academic performance.

The results consistently support the conclusion that integrating technology into formative assessment has a statistically significant and positive impact on learners' performance. The consistently significant negative mean differences observed in the data indicate that the experimental groups, which received the technological intervention, outperformed the control groups by a considerable margin. These findings strongly reinforce the initial hypothesis that technology-enhanced formative assessment improves learning outcomes. Moreover, the outcomes of this study offer compelling implications for pedagogical practice by highlighting the potential of technological tools to enhance the effectiveness of formative assessment, thereby informing future decisions regarding instructional design and assessment strategies in EFL contexts.

4. DISCUSSION

In contemporary language education, the integration of technology into instructional practices has garnered increasing attention, particularly in relation to enhancing the efficacy of formative assessment. This growing interest has provided the impetus for the present study, which aims to offer a comprehensive and empirically grounded exploration of the potential of technology to transform formative assessment practices in English as a Foreign Language (EFL) context. Specifically, the study investigates the central research question: *To what extent do technology-oriented practices optimize EFL learners' performance in formative assessment?*

By addressing this critical inquiry, the study seeks not only to evaluate the pedagogical value of technological tools in formative settings but also to contribute to improving the educational experiences and learning outcomes of EFL students. As demonstrated by the empirical data and statistical analyses discussed in the preceding sections, the findings underscore the transformative potential of technology in fostering meaningful assessment-driven learning. These results provide a strong foundation for rethinking conventional assessment approaches and advocate for a more dynamic, evidence-based integration of digital tools in language teaching and learning environments.

The findings of this study reveal a significant enhancement in the performance of the experimental groups that were exposed to technological intervention, compared to the control groups that relied on traditional instructional methods. This notable disparity underscores the effectiveness of integrating educational technology into formative assessment practices. Within this framework, Bhagat and Spector (2017) emphasized the necessity of conducting further research on the role of technology in supporting formative assessment. Their work highlighted the potential of technological tools not only to improve academic performance but also to positively influence learners' attitudes and motivation—an outcome echoed in the present study.

Similarly, Elmahdi, Al-Hattami, and Fawzi (2018) demonstrated that the integration of technological tools enhances both teaching efficiency and student learning outcomes by increasing engagement, optimizing instructional time, and fostering a more enjoyable learning environment. Additionally, Warschauer and Healey (1998) noted that technology supports interactive learning environments, which are essential to promoting language acquisition and overall academic performance. Collectively, these studies affirm the relevance and significance of the current research, reinforcing the broader scholarly consensus on the pedagogical benefits of technology-enhanced formative assessment. Most importantly, this study contributes a focused, data-driven examination of how such integration directly optimizes EFL learners' performance—marking a substantive addition to the existing body of literature.

In the pursuit of rigorously measuring the impact of technological intervention on EFL learners' formative assessment performance, the current study is firmly anchored in Kane's (1992, 2006, 2013) argument-based approach to validity. This framework serves as a foundational pillar for evaluating the strength and coherence of the inferences drawn from assessment outcomes. It provides a comprehensive structure for ensuring that the conclusions derived from empirical data are both credible and defensible. Central to this framework is the emphasis on validating interpretations and uses of test scores through a clearly articulated chain of reasoning supported by empirical evidence. In the context of this study, Kane's model is applied to substantiate the claim that technology-enhanced formative assessment significantly improves learner outcomes, while also acknowledging and addressing potential counterarguments regarding the efficacy of such interventions.

A critical component within this framework is the validity of scoring inferences. These inferences are instrumental in establishing the reliability and interpretability of assessment scores. They ensure that the data accurately reflect student performance and that any conclusions drawn are not only statistically sound but also educationally meaningful. Accurate scoring enables educators and researchers to confidently attribute observed changes in performance to the intervention itself—namely, the integration of technology—rather than to extraneous variables or measurement error. Therefore, adhering to Kane's model allows for a systematic validation of each stage of the assessment process, from scoring and generalization to extrapolation and implications, reinforcing the study's claim that technological tools can elevate the efficacy of formative assessment in EFL contexts.

The foundational assumption underpinning this study is that scoring methods were applied consistently across both the experimental and control groups, thereby ensuring that any observed differences in learner performance can be attributed to the technological intervention rather than discrepancies in assessment procedures. To uphold this standard of fairness and internal validity, a standardized scoring rubric was implemented across all assessments. Each correct answer was awarded one point, and responses containing spelling errors were automatically disqualified in the vocabulary and grammar sections, given the study's primary focus on language form. However, a nuanced exception was made for the final four items in

the "functions" section of the assessment, where a half-point was awarded if the response contained minor spelling errors. This decision aligns with the communicative nature of functional language use, which is often evaluated for meaning and intent rather than strict linguistic accuracy.

By codifying these scoring criteria in advance and applying them uniformly, the study eliminates the possibility of scorer bias or inconsistencies that could otherwise compromise the validity of the results. Furthermore, these procedures mitigate concerns regarding the influence of raters' awareness of group assignments on scoring decisions. This rigor in scoring strengthens the argument that the improved outcomes observed in the experimental groups are genuinely attributable to the technological intervention rather than to any extraneous variables related to evaluation practices.

Moreover, the use of technological intervention in assessment enhances the objectivity of evaluating learners' performance, thereby reducing the influence of human error and subjective judgment. This assertion is grounded in the recognition that various external factors—such as assessor fatigue, mood, or perception biases—can unintentionally alter the outcome of traditional assessments. For instance, paper-based assessments frequently pose challenges related to handwriting legibility, which may inadvertently influence scorers to assess aspects unrelated to the intended construct. Additionally, test anxiety is a commonly documented phenomenon in conventional assessment settings, and it has been shown to negatively affect learners' performance, thus compromising score validity.

In contrast, the integration of technology—specifically through platforms such as Plickers—promoted a more engaging and anxiety-reduced testing environment. Learners exhibited increased motivation and enthusiasm, resulting in more accurate representations of their actual competencies. As Beatty (2010) noted, gamified and interactive digital assessment tools in EFL contexts can foster heightened student participation and positively influence performance outcomes. While one might argue that such technological tools could introduce new forms of bias—such as advantages for more tech-savvy learners—this concern was preemptively addressed in the present study. All participants received uniform instruction on how to use the Plickers system, including guided demonstrations and comprehension checks, thus minimizing disparities in digital familiarity and ensuring equitable conditions across the experimental cohort.

A further key scoring inference lies in the alignment of assessment tasks with the instructional objectives across both the control and experimental groups. In this study, all participants were exposed to identical curricular content, guided by the same learning competencies, objectives, and instructional units. This parity ensures that the assessment scores authentically reflect the intended language competencies, minimizing the risk of construct-irrelevant variance. Any assertion that the technological intervention introduced fundamentally different task types or assessment formats is unfounded, as both groups completed equivalent tasks derived from the same assessment instruments. Thus, the differentiation in performance outcomes can be credibly attributed not to content variation but to the affordances provided by the technological tools integrated into the experimental condition.

These technological variables—such as the use of visual supports, immediate feedback, diverse response formats, increased learner engagement, reduced anxiety, and enhanced teacher-learner interaction—constitute the core mechanisms driving the improved outcomes observed in the experimental group. Such elements are widely acknowledged in the literature as critical contributors to optimizing language learning environments (Elmahdi et al., 2018; Beatty, 2010). Consequently, the observed disparity in post-test performance between the groups can be confidently ascribed to the pedagogical and cognitive advantages enabled by the

technological intervention rather than any discrepancy in assessment design or instructional content.

Another critical scoring inference pertains to the role of technology in delivering immediate feedback, a feature that significantly contributes to enhancing learners' performance during formative assessment, as evidenced in their improved scores. Immediacy—recognized as a salient technological variable—facilitates the prompt recognition and correction of errors, reinforcing learning through real-time reflection and adjustment. Bhagat and Spector (2017) emphasized that delays in feedback delivery can hinder student engagement and learning efficacy. They further argued that when feedback is not provided promptly, its constructive value may be diminished, potentially undermining its intended instructional purpose. This positions immediacy as one of the most compelling advantages of integrating technological tools in assessment contexts (Dakka, 2015; Robertson et al., 2019).

In support of this claim, Marsh, Lozito, Umanath, Bjork, and Bjork (2012) demonstrated that verification feedback administered immediately after each test item resulted in significantly higher performance compared to delayed feedback, such as the post-assessment distribution of answer keys. Given that immediacy is inherently a technological affordance, any critique suggesting that it unfairly benefits experimental groups overlooks the fact that it constitutes an intentional design variable under investigation, rather than an uncontrolled confound. A further inference involves the capacity of technological platforms to standardize testing conditions, thereby minimizing external variables that might otherwise influence performance. Although technical issues such as connectivity problems or platform unfamiliarity may arise, these risks were proactively mitigated in the current study through pre-assessment readiness checks and comprehensive instruction sessions. This ensured that all participants had equitable access and understanding of the technological tools prior to assessment administration, thus preserving the reliability and validity of the resulting data.

Shifting toward generalization and extrapolation, several key inferences are essential to extend the findings of this study and inform both pedagogical practices and educational policy. A central assumption is that the statistically significant performance improvements observed in the experimental groups suggest that integrating educational technology can effectively enhance formative assessment outcomes among English as a Foreign Language (EFL) learners across a wide spectrum of proficiency levels—from beginners to advanced students. While it could be argued that varying levels of learners' technological literacy might contribute to disparities in outcomes, this concern is mitigated by the reality that students today are immersed in a digitally saturated environment. Consequently, there is a pressing need for both teachers and learners to receive targeted training on how to effectively utilize a diverse range of digital tools and platforms that align with instructional goals and learning needs. With appropriate training and support, similar improvements in formative assessment performance are likely to be replicated in other educational contexts.

Regarding the geographical extrapolation of these results, the notable success of the technological intervention in the current study suggests the potential for similar outcomes in varied regional and institutional settings. However, it is important to acknowledge the contextual challenges that may influence implementation. In particular, rural schools may face infrastructural constraints, limited access to technological tools, or insufficient professional development opportunities for educators. Chapelle (2011) draws attention to such obstacles, identifying inadequate infrastructure, resistance to change, and lack of teacher training as key barriers to the effective integration of educational technology. These challenges, while significant, should not deter innovation; rather, they underscore the need for systemic support

and strategic policymaking to ensure equitable access to the benefits of technology-enhanced formative assessment across all learning environments.

Drawing from the findings of this research, several pedagogical and policy-level implications emerge. Foremost, the integration of educational technology into formative assessment highlights a critical need for strategic investment by policymakers in both infrastructure and continuous teacher training. This includes the allocation of substantial financial resources and the provision of ongoing technical support and maintenance to ensure sustainability. Stockwell (2013) emphasized that the consistent use of mobile-assisted language learning tools has a long-lasting positive effect on students' language skills, thereby underscoring the value of long-term technological integration.

Moreover, the study infers that technology-based formative assessment can foster greater student engagement and motivation through interactive and visually enhanced modalities. These dynamic features serve as powerful catalysts for learning across a variety of educational settings. Although individual learning preferences and cognitive styles may influence the degree of receptiveness to technological tools, such variability can be accommodated through the flexibility and adaptability that digital platforms inherently offer. Thus, these tools not only support differentiated instruction but also encourage learner autonomy and responsiveness.

In addition, the study opens avenues for future research focused on exploring and refining additional technological tools and strategies to enhance formative assessment. This direction does not suggest the displacement of conventional assessment methodologies; rather, it encourages a complementary approach where different strategies are adopted based on effectiveness and context-specific needs. A further implication posits that technological innovations may serve as a foundational pillar in broader educational reforms. The ability of educational technology to transcend traditional assessment boundaries situates it as a transformative force in contemporary learning environments.

Nonetheless, it is important to recognize potential counterarguments. One such concern is that an overemphasis on technology might overshadow essential aspects of education, including curriculum development, pedagogical approaches, and meaningful teacher-student interactions. However, this concern is mitigated by the understanding that technological implementation should not replace but rather enhance and integrate with existing educational practices. As long as this integration is balanced and pedagogically grounded, the potential benefits of technology in formative assessment and beyond remain both valid and impactful.

While several limitations have been addressed through counterarguments within the analysis, it is essential to explicitly underscore some critical shortcomings that warrant further consideration. First, this study employed a quantitative research design aimed at providing empirical evidence to support or reject the directional hypothesis concerning the effectiveness of technology in formative assessment. Although this approach offers statistical rigor and objectivity, it lacks the nuanced insights that a qualitative perspective could provide. Future research is therefore strongly encouraged to adopt qualitative or mixed-methods designs to explore learner and teacher experiences, attitudes, and contextual factors that underlie the observed quantitative outcomes. Such an approach would offer a richer understanding and contribute more holistically to the existing body of literature.

Another notable limitation pertains to the sample size. Each group in the two experiments consisted of relatively small cohorts (ranging from 24 to 29 participants), which may affect the generalizability of the findings. Small sample sizes can introduce sampling variability and may not fully capture the diversity of learner profiles present in larger

educational populations. While the study's design—incorporating two separate experiments—serves as a mitigating strategy to enhance reliability and internal validity, caution should be exercised in extending the findings to broader contexts without replication on a larger scale. Future research would benefit from incorporating larger, more diverse samples across multiple institutions and geographic locations to enhance external validity and applicability.

5. CONCLUSION

This study critically examined the impact of integrating educational technology on enhancing EFL learners' performance within the scope of formative assessment. Anchored in a robust theoretical foundation, it drew upon literature covering formative assessment theory, Kane's framework of argumentative validity, and the evolving role of technology in instructional assessment. The findings substantiate a compelling case for the efficacy of technology-enhanced formative assessment, revealing statistically significant gains among learners exposed to the intervention compared to those assessed through conventional means.

Framed by Kane's model, the analysis demonstrated that technology integration not only reinforces theoretical validity through improved scoring, generalization, and extrapolation inferences but also yields measurable pedagogical benefits. Specifically, the use of platforms such as Plickers enabled the delivery of timely, personalized feedback, heightened learner engagement, and facilitated instructional responsiveness—all of which contributed to improved academic performance.

A notable contribution of this research lies in its practical implementation of Plickers as a formative assessment tool within the EFL context, offering a replicable model for educators and institutions. Given the exponential growth of educational technology—with over 389,000 learning applications reported across digital platforms in 2023—the findings underscore the vast potential for further exploration and innovation in this domain. While the study addresses key limitations and responds to potential counterarguments, it ultimately calls for broader, cross-contextual research into technology integration. Such inquiry is essential to guide future policy decisions and pedagogical reforms aimed at leveraging digital tools to enhance language learning outcomes across diverse educational landscapes.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37*, 1–15.

Andrade, H., & Heritage, M. (2017). *Using formative assessment to enhance learning, achievement, and academic self-regulation*. Routledge. https://doi.org/10.4324/9781315623856

Angoff, W. H. (1988). Validity: An evolving concept. *Educational Measurement: Issues and Practice*, *7*(1), 19–22.

Beatty, I. D., & Gerace, W. J. (2009). Technology-enhanced formative assessment: A research-based pedagogy for teaching science with classroom response technology. *Journal*

of Science Education and Technology, *18*(2), 146–162. https://doi.org/10.1007/s10956-008-9140-4

Beatty, K. (2010). *Teaching and researching computer-assisted language learning* (2nd ed.). Routledge.

Bell, B., & Cowie, B. (2000). The characteristics of formative assessment in science education. *Science Education*, *85*, 536–553.

Bhagat, K., & Spector, J. (2017). Formative assessment in complex problem-solving domains: The emerging role of assessment technologies. *Educational Technology & Society*, *20*(4), 312–317.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–74.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. Educational Assessment Evaluation and Accountability. 21. 5-31. 10.1007/s11092-008-9068-5.

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551–575.

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.), *Educational evaluation: New roles, new means* (pp. 26–50). University of Chicago Press.

Bloom, B. S., Madaus, G. F., & Hastings, J. T. (1971). *Handbook on formative and summative evaluation of student learning*. McGraw-Hill.

Bonk, C. J., & Graham, C. R. (Eds.). (2006). *Handbook of blended learning: Global perspectives, local designs* (pp. 8–10). Pfeiffer Publishing.

Broadfoot, P., Weeden, P., & Winter, J. (2002). *Assessment: What's in it for schools?* Routledge. https://doi.org/10.4324/9780203468920

Brookhart, S. M. (2007). Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 43–62). Teachers College Press.

Caldwell, J. E. (2007). Clickers in the large classroom: Current research and best-practice tips. *CBE-Life Sciences Education*, *6*(1), 9–20.

Chapelle, C. A. (2011). *Computer applications in second language acquisition: Foundations for teaching, testing, and research* (2nd ed.). Cambridge University Press.

Chazan, D., Olsher, S., & Yerushalmy, M. (2016). How might the use of technology in formative assessment support changes in mathematics teaching? *For the Learning of Mathematics*, *36*(3), 11–18.

Creswell, J. W. (2019). Research Design: Qualitative, Quantitative, and Mixed Method Approaches. Sage Publications.

Cronbach, L. J. (1971). Test validation. In R. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.

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

Dakka, S. (2015). Using Socrative to enhance in-class student engagement and collaboration. *International Journal on Integrating Technology in Education*, *4*(3), 13–19. https://doi.org/10.5121/ijite.2015.4302

Dalby, D., & Swan, M. (2018). Using digital technology to enhance formative assessment in mathematics classrooms. British Journal of Educational Technology. https://doi.org/10.1111/bjet.12606

Danielson, C. (2011). *Enhancing professional practice: A framework for teaching* (2nd ed.). ASCD.

Elmahdi, I., Al-Hattami, A., & Fawzi, H. (2018). Using technology for formative assessment to improve students' learning. *The Turkish Online Journal of Educational Technology*, *17*(2), 182–188.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). American Council on Education.

Foxcroft, C. D., & Roodt, G. (2013). *Introduction to psychological assessment in the South African context* (4th ed.). Oxford University Press.

Garrett, H. E., & Woodworth, R. S. (1973). *Statistics in psychology and education*. Vakils, Feffer and Simons Private Ltd.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, *6*(4), 427–438.

Habler, B., Major, L., Warwick, P., Watson, S., Hennessy, S., & Nicholl, B. (2016). *Perspectives on technology, resources and learning: Productive classroom practices, effective teacher professional development*. University of Cambridge.

Haeffel, G., & Cobb, W. (2022). Tests of generalizability can diversify psychology and improve theories. *Nature Reviews Psychology*, *1*(4), 243–244. https://doi.org/10.1038/s44159-022-00039-x

Haughney, K., Wakeman, S., & Hart, L. (2020). Quality of feedback in higher education: A review of literature. *Education Sciences*, *10*(3), 60. https://doi.org/10.3390/educsci10030060

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Corwin Press.

Herold, B. (2016). *Technology in education: An overview*. *Education Week*. Retrieved from https://www.edweek.org/ew/issues/technology-in-education/

Irving, K. (2015). Technology-assisted formative assessment. In *Encyclopedia of information science and technology* (3rd ed., pp. 1738–1746). IGI Global. https://doi.org/10.4018/978-1-4666-9616-7.ch017

Jamieson, J., & Musumeci, M. (2017). Integrating assessment with instruction through technology. In C. A. Chapelle & S. Sauro (Eds.), *The handbook of technology and second language teaching and learning* (pp. 320–334). Wiley. https://doi.org/10.1002/9781118914069.ch20

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535.

Kane, M. T. (2006). Validation. In R. B. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.

Landis, J.R. and Koch, G.G. (1977) An Application of Hierarchical Kappa-Type Statistics in the Assessment of Majority Agreement among Multiple Observers. Biometrics, 33, 363-374. https://doi.org/10.2307/2529786

Leedy, P. D., & Ormrod, J. E. (2010). *Practical research: Planning and design* (9th ed.). Pearson Education.

López-Pastor, V., & Sicilia-Camacho, A. (2017). Formative and shared assessment in higher education. Lessons learned and challenges for the future. *Assessment & Evaluation in Higher Education, 42*(1), 77–97.

Mäkipää, T., & Hildén, R. (2021). What kind of feedback is perceived as encouraging by Finnish general upper secondary school students? *Education Sciences*, *11*(1), 12. https://doi.org/10.3390/educsci11010012

Marsh, E. J., Lozito, J. P., Umanath, S., Bjork, E. L., & Bjork, R. A. (2012). Using verification feedback to correct errors made on a multiple-choice test. *Memory*, *20*(6), 645–653.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2010). *Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies*. U.S. Department of Education.

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

Nitko, A. J. (2004). *Educational assessment of students* (4th ed.). Pearson Education.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Popham, W. J. (2006). *Assessment for educational leaders*. Allyn & Bacon.

Popham, W. J. (2008). *Transformative assessment*. ASCD.

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, *46*(4), 265–273.

Rahi, S. (2017). Research design and methods: A systematic review of research paradigms, sampling issues and instruments development. *International Journal of Economics and Management Sciences*, *6*(2), 1–5.

Robertson, S., Humphrey, S., & Steele, J. (2019). Using technology tools for formative assessments. *The Journal of Educators Online*, *16*(2), 1–15. https://doi.org/10.9743/JEO.2019.16.2.11

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Rand McNally.

Shaw, J., & Crisp, R. J. (2011). *Validity in educational and psychological assessment*. SAGE Publications.

Spector, J., & Yuen, A. (2016). *Educational technology program and project evaluation*. Routledge. https://doi.org/10.4324/9781315724140

Stiggins, R. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*(10), 758–765. https://doi.org/10.1177/003172170208301010

Stockwell, G., & Hubbard, P. (2013). *Some emerging principles for mobile-assisted language learning*. The International Research Foundation for English Language Education.

Urbina, S. (2014). *Essentials of psychological testing* (2nd ed.). Wiley. https://doi.org/10.1002/9781394259458

Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching*, *31*(2), 57–71.

Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment*, *11*(3&4), 283–289. https://doi.org/10.1207/s15326977ea1103&4_7

Wiliam, D., & Leahy, S. (2007). A theoretical foundation for formative assessment. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 29–42). Teachers College Press.

Wiliam, D. and Thompson, M. (2008) Integrating Assessment with Instruction: What Will It Take to Make It Work? In: Dwyer, C.A., Ed., The Future of Assessment: Shaping Teaching and Learning, Lawrence Erlbaum Associates, Mahwah, 53-82.