Social Sciences
2025, Vol. 14, No. 1, pp. 8-21
https://doi.org/10.11648/j.ss.20251401.12

Research Article

An Empirical Assessment of Moroccan EFL Teachers' Use of Generative AI for EFL Formal Assessment: An Intervention Study

Far-hat Maryam*, Ouchouid Jamaa

Department of English Studies, Caddy Ayyad University, Marrakech, Morocco

Abstract

This paper offers an empirical analytical assessment of the use of a generative Artificial Intelligence (AI) chat box tool – ChatGPT- by Moroccan EFL teachers in designing formal tests. The field of language teaching and learning is experiencing an unprecedented pedagogical influence with the inception of AI. This machine-guided program has made easier multitude of pre-lesson and in-classroom practices such as assessment. The latter is the 'ongoing' process of competence-based and performance-based evaluation that situates a learner at a certain knowledge level. One type of assessment is 'formal assessment' which is any form of acquired knowledge evaluation that is pre-planned rather than prompted (ibid, pp. 5-6). Giving the recency of AI in education, a growing need appears for guidance into the art of its use. Accordingly, this paper aims to provide a comprehensive survey of how ChatGPT is used by Moroccan EFL teachers in designing formal tests. The researcher sheds light on the content validity of teachers' prompts and the extent to which they target the mélange of testing principles and objectives. The paper includes a pre-intervention stage where the teachers' prompts to ChatGPT to generate a test. The intervention is based on participants' open-ended prompts and constitutes ample practical tips to illustrate the valid use of ChatGPT. The post-intervention assesses the teachers' aspired development. The findings revealed that enabling Moroccan EFL teachers with a guide including prompt designing and test construction principles positively influenced their performance in using ChatGPT to generate formal tests. It was recommended, among others, that AI use in teaching and planning be integrated in their professional training.

Keywords

AI in Education, ChatGPT, EFL Assessment, Formal Tests

1. Introduction

The present era is experiencing profuse change with the inception of AI. The monopoly of AI is a de facto recognition given its partaking in many areas of human interest and function. Sic, aspects of human performance and competence have been and need to be altered in response to this accelerated change. Among the fields of AI interference are English language teaching and learning. For instance, the fulfilled Bachman's English language competence has become a heated necessity given the language popularity over the globe. Investigations and implementations into the pedagogy and

*Corresponding author: m.farhat.ced@uca.ac.ma (Far-hat Maryam)

Received: 13 December 2024; **Accepted:** 8 January 2025; **Published:** 24 January 2025

curriculum of English teaching and learning is witnessing currently not just the interreference of human brain but also that of the programmed machine 'brain'; i.e., AI. This intervention is getting away with old-fashioned methods of planning, management, and assessment. Also, it helps the journey to be less time and energy-consuming, accessible, and innovative for all the stakeholders in language education; namely, for teachers. Currently, many generative AI tools pop to provide comprehensive guidance for teachers such as Chat GPT- a generative AI tool for assisting teachers in designing lesson plans and assessment based on prompts and explanations.

Language teaching and learning is one domain of AI convergence which needs reflective observations, plans, and actions to regulate its integration. Improving on language education equals monitoring how AI can be used by both the traditional knowledge providers and receivers. Among the frames of language teaching and learning largely touched upon by AI is student's 'assessment'. Learners are destined to receiving a high dose of assessment. Assessment is one popular subject of pedagogy and is an observatory and analytical process of decision-making concerning a language knowledge level of learners. This era has witnessed a transition from traditionally made assessment to an array of AI generated assessment directing the process from methodology to post-assessment action-taking. On one hand, this artificial assistant has opened new options for assessors giving them a variety of formal testing techniques and informal assessment instructions. On the other hand, the integration of AI in language assessment has put to the front questions and concerns such as the reliability of teachers' prompts given that each user has their own modus operandi to which the machine would respond in a given way.

The present paper will touch on the use of ChatGPT by Moroccan EFL teachers to generate tests. The fact that this AI tool requires no skill or prior training for use makes teachers assume the correctness of use and, likewise, makes researchers inconsiderate of the possibility of incorrectness of use. The purpose is to examine the 'correctness' of teachers' in terms of proper prompt-designing and proper use of the prompt to design an assessment that has the necessary features.

This study is an empirical one. It begins with a hypothesis that Moroccan EFL teachers' prompts to ChatGPT to generate an assessment need optimization within intervention. It is an assessment given that it assesses teachers' prompts in search of weakness to improve on their performance. The study goes through three stages: a pre-intervention, a while-intervention, and a post intervention. It begins by a questionnaire to prove the hypothesis. Data gathered from this questionnaire are used to initiate an action. The latter is a guide provided to teachers that targets two major areas: general proper prompt-designing ("Prompt design strategies," n.d.) and designing a prompt that respects Brown test construction principles [15]. To assess the post-intervention performance, the same respondents are given a questionnaire to examine their new prompts.

2. Preliminaries

2.1. Assessment

The gift of knowledge is the foundation of mind prosperity. For prosperity to grow affluently, self or others' reflection is the premise along with decision-making and action-taking procedures. Accordingly, every act of learning necessitates some evaluation and judgment of its evolvement. Language learning is not different from this eternal guideline. Constructive judgment of language learning has been established for decades so as to ensure the positivism of this process. Using the 'dynamic pedagogy 'thought, this judgment is known as 'assessment'. Assessment is directed towards situating learners in a certain level of achievement using a mélange of evidence gathering techniques such as objective-directed observation, reflective questioning, and graded web-based or paper-based testing procedures.

Assessment is an interdependent component of the learning process given both intervenes in and is determined by other related pedagogical aspects such as instruction and planning. It provides insights into the quality of both language teaching and learning. Assessment in ELT is a systematic process of collecting, evaluating, and making use of information about the language knowledge level of learners. It helps evaluate the outcomes of learning and design and implement future instructions. Brown views assessment as incorporating a variety of activities, such as testing and evaluation [15]. All the activities seek to provide an understanding of learners' abilities.

The basic premise of assessment is anything diverging from the frame of interpretivism. Assessment should stand for knowledge enhancement and prosperity. It is intended to help assessors in planning future paths into knowledge diffusion. In ELT, assessment overlaps with two other terms: evaluation and testing. While assessing is an ongoing procedure of performance measuring, testing is an assessment technique. As for evaluation, it encompasses the two terms given that the evaluator could resort to other data collection techniques such as observation and self-reflection.

Assessment is pivotal for several reasons including the utility of feedback in enhancing teaching methods and activities. Also, it helps to detect the strengths and weaknesses. Besides, assessment can help teachers know the learning progress and make decisions about teaching strategies and learning goals following Drucker's 1954 quote: "If you can't measure it, you can't improve it.". Also, assessment in ELT has shifted the focus from linguistic competency to communicative competence given a multitude of outcomes.

2.1.1. Types of Assessment

What inevitably comes to the mind when the term assessment is used is the traditional paper-based test that teachers

administer in a classroom environment where performance is equaled to a certain grade on the sheet. However, there are diversifying techniques of assessment that could take the form of mere observation or shortened feedback. Different assessment forms and techniques target different element of learning that basically related to competence and performance. Simultaneously, assessment is the first option for teachers to spot the weaknesses in their students' body of knowledge before, during and after the learning process.

Assessment can be classified into different types given their construction and instruction. The most common dichotomy is 'formative' vs. 'summative' assessment. According to Black & Wiliam [14], 'formative' assessment is a process that helps teachers identify the progress of their learners and adapt their teaching strategies to suit it. It involves an ongoing observation, evaluation, and feedback. Brown [15] defines formative assessment as taking place during the course and intended towards enhancing and adapting teaching choices based on ongoing feedback whereas summative assessment is instructed at the end of a course to rate students' achievement. Gronlund & Waugh [16] explain that summative assessment is a standardized test form for evaluating students' achievement that usually is administered at the end of a course. 'Summative' assessment can be opposed to 'diagnostic' assessment which is executed at the beginning of a course. The objective here is to detect the strengths, and weaknesses of the testees. It informs instructional planning by helping teachers tailor lessons to meet learners' needs [19]. The second distinction can be made between 'formal' and 'informal' assessment. While formal assessment is "systematic, planned sampling techniques constructed to give teacher and student an appraisal of students achievement." and is executed using written tests or projects [15], informal assessment is unplanned and subjective observation and reflection "designed to elicit performance without recording results and making fixed judgments about a student's competence." [15]. The difference can also be made between 'objective' and 'subjective' assessment. As the label implies, objective assessment is reliable in that two graders would have the same grade for it such as a multiple-choice test. Subjective assessment involves other testing techniques such as composition where the grader could use personal judgment on its quality.

Assessment can be conducted using varied techniques such as mere observation where teachers observe the language use of leaners in different classroom contexts. Tests and Quizzes: Standardized tests, quizzes, and exams evaluate learners' knowledge of grammar, vocabulary, reading, and writing skills. These assessments provide measurable outcomes for future instructions. Portfolios, compilations of students' work during a course which show their achievements, are also among the assessment techniques [17]. Harmer [20] specifies different types of tests such as 'discrete items'. They are intended to test a specific objective at a time. 'Direct test item' is when students are tested in their ability to do things with language such as perform a dialogue or compose a piece of writing. The 'indirect test items' refer to the testing techniques that usually target one element such as gap filling in and matching.

2.1.2. Principles

The process of assessment follows established guidelines to be prepared and processed adequately. Harmer [20] sets three criteria for a good test: validity, reliability, and washback effect. A test is valid when it tests what it intends to assess. This criterion guarantees that the results of assessment mirror testees' true abilities [18]. 'Reliability' is where testees and correctors would have an equal opportunity to give answers and score them, respectfully. Bachman & Palmer [13] note that a test is reliable when it gives the same results with the same tested group and conditions. The third criterion is the 'washback effect': when teachers only adapt their teaching for the test. Brown specifies another criterion of good test; 'authenticity'. A test is authentic when it uses materials and activities that reflect real-life meaningful contexts. Taxonomies guiding the objective and process of assessment diversify. Bloom's 1957 taxonomy is one popular framework showing the element that need testing at a certain learning level. This framework classifies the cognitive learning process within a hierarchy of targeted skills starting from information remembering to information creating. Thus, students should not be tested at an upper level in developed skills such as analyzing theories and creating realms of though while they are still in lower levels of information reception, perception, and mere production.

2.2. AI in Education

2.2.1. Definition of AI

The term AI is a recency in the universal dictionary of vocabulary. It is short for 'Artificial Intelligence'. It is 'artificial' given its nonhuman compute-based state of being. The 'intelligence' component stems from its capacity to endlessly generate ready knowledge for the prompt producer. Knowledge itself varies from pure episteme and reasoning to know-how procedural knowledge. The philosophy behind the creation of AI was to serve two main aims. The first aim is to help humans perform complex tasks with less effort and devoted time. AI is intended to target human energy to other tasks requiring bare human intervention. This reflects the 'assistance' feature of the software. The second aim is to 'simulate' humans in operating. This means that not just they help humans in doing tasks but they stand as 'them' in doing the tasks.

The thriving of humans over cognitive human-like entities has always been existing. Buchanan [2] illustrates the story and timeline of the birth of AI. Accordingly, he traces back the term 'intelligence' to Greek mythology with Hephaestus and Pygmalion who first introduced the idea of human-like entities. Aristotle' syllogistic logic was the basis for establishing

the practical knowledge of AI. The idea was then first embodied in the 13th century with what Buchanan [2] refers to as 'talking heads'- audio-visual head-like machines- [2]. The creation of machines simulating a certain human skill developed through centuries. The very known invention is that of printing in the 15th Century, clocks in the 16th, and primitive calculators in the 17th. Among the 19th century scientists establishing algebraic and mechanical fundamentals of AI are George Boole and Charles Babbage & Ada Byron. With the invention of certain basic machines in the beginning of the 1900s, the question of whether humans can simulate the human though became louder. This idea was translated to reality by, first, a 1950 paper by the British polymath Alan Turing entitled 'Computing Machinery and Intelligence' where he explored the possibility of constructing machines with human intelligence. It first begun with the conceptional frameworks shaping the basis robotics and machinery by scientists like Alan Turing back in 1947. The latter invented the 'Turing test' which assigns the intelligence quality to machines based on their ability to mimic some human cognitive capacities [6]. Different contributions made the idea clearer over time to the extent that in the second half of the 20th century the term "artificial intelligence" was coined by the computer scientist John McCarthy in 1956 who defines it as "the science and engineering of making intelligent machines, especially intelligent computer programs." [7]. The quest reached its revolutionary turn with the invention of Strachey's checkers program in 1951 by the British computer scientist and semanticist Christopher S. Strachey. This program intends to play a quick and complete human-like game of draughts [8]. This initiated the trend of machine programming which was manifested in other 20th century machines and programs such as 'Logic Theorist' (1956) and the 'General Problem Solver' (1957). Thus, AI was created by an accumulation of non-computer machines and programming computers.

2.2.2. The Use of AI in Education

Two tenets are said to be behind the invention of AI: assisting humans and imitating humans. The assisting quality is about providing guidance into doing a task for the human. As for the imitation, AI intends to do the task for humans simulating their cognitive and reasoning processing in going through tasks. McCarthy [7] admits that "the ultimate effort is to make computer programs that can solve problems and achieve goals in the world as well as humans". Among the fields of AI intervention is education. This applicability combines both assistance and imitation features. AI can help teachers in pedagogy and its interconnected aspects with both curriculum and assessment. Among the latest largely used generative AI media are MagicSchool- an umbrella AI model that combines different tools for, for instance, planning, assessment, and material adaptation-, Formative AI which helps in creating formative assessment samples, and Canva- an AI tool for designing creative slides for any suggested topic. However, the most sophisticated deep learning tool is ChatGPT. It is sophisticated given its advanced capabilities of generating human-like prompts in a machine-like way. It is a deep learning tool given its flexible programming features that enables it to learn and enhance its performance from the data available either from AI itself or from its conversational interaction with users. ChatGPT is generally known to be the 'the golden child' of AI [5]. It is the abbreviated form of "Generative Pre-trained Transformer" [3]. This AI bot is the first escape for humans needing informative answers for questions in different fields, varied texts generation, interactive engaging. In the field of language education, this bot is largely used by teachers.

All those AI bots assist teachers with customized resources covering mainly the pre-classroom preparation phase and partly while and post classroom execution, grading, and reflection. This would save the teachers the time and energy they would instead devote in the in-classroom implementation of those resources. For instance, AI bots could be used by teachers to generate lesson plans covering a set of learning and teaching objectives. All that a teacher would do for example is use MagicSchool to design a lesson by entering a prompt including specifics of the lesson such as the targeted level, its learning styles, the integrated skills, the type of teaching materials, and the related curriculum themes and objectives. Most other AI bots would assist in lesson planning in either providing detailed guides, or presenting the content in a ready-to-use way. Such those bots are Canva. The latter is largely used not only by educators but by any who is in search of arresting informative data as a rapid time. Another intervention would be using AI for material adaptation. The bot would help adjust the content of the reading or listening material in accordance with the intelligence and learning style of leaners and their language knowledge level. Also, the AI bot would properly blend the material in a lesson plan stage and provide instruction for executing it with respect the classroom language. A one popular AI tool for material adaptation is Diffit which both levels resources and generates suitable ones. Among the AI tools for grading are Gradescope. The latter is suitable for adding grading subjectivity by providing a set of grading criteria and feedback instructions for teachers. The AI bots can assist in a variety of realms such as 'ideas generation'. It can help "generate texts in a very short amount of time, making it easier and more efficient to search for and find summarized information and ideas related to the subject of interest" [4]. Moreover, the AI bots can help educators design formative tests addressing learning needs and testing objectives by specifying in the prompt the specifics of the test and the tested learners. ChatGPT is the most known multifaceted AI bot in the enhancement of EFL teaching and learning experience. It can assist in lesson planning by helping in ideas generation, resources providing, instructions designing, assessment procedures and grading criteria [10]. All that an educator need to do is to direct the AI bot using the correct prompt.

Generally, since the appearance of generative AI, its im-

plementation in education has attracted heated investigations in the past years. The investigations namely attract three areas. The first of which is how it can be used accurately to cater for many de facto pedagogical requirements. Indeed, the recency of generative AI makes its use in education a novelty that is yet unknown to a large group of educators. As for those who have had the experience, the expected fact would be their yet unskilled interaction with the AI tool. For this, there are different guides into the 'proper' use of AI bots so that the purpose behind the use can be accurately achieved. A skilled user of AI would thus be able to use it correctly to target a certain pedagogical aspect such as designing formative assessment tests. The second area has to do with how it can be used carefully in fulfilling the requirements. This, in a way, questions the ethical aspects of using AI that relate to the originality of the work done. The third are is how it can be developed in a way that correlates between the first and second tenets. This has to do, first, with deep learning that would enable generative bots to improve on their responses base on their interactions with the prompt writer and, second, with the intervention of programmers to enhance on other aspects such as its ethical side. For indeed, given this eyeblink spread of AI use in education, educators should be endowed with a basic comprehensive understanding of AI to engage positively and productively with this technology.

2.2.3. The Use of AI in the ELT Moroccan Context

The Moroccan EFL context has not been safe from the invasion of AI. Sic, Moroccan EFL teachers on different levels implement AI bots in their pedagogical decisions and construction. Since the vast appearance of online classes during the coronavirus pandemic, the E-trend has become a preference of educators and students likewise. Remote education has exploited digital platforms to connect teachers and students providing interactional lessons, online assessment, and feedback. This remote education unleashed an advanced version of technology with the adoption of AI. The latter is currently profusely and gradually used in the Moroccan EFL context. Its platforms are aspired to provide easy access to information beyond the traditional classroom context. For instance, the Moroccan higher education sector is implementing AI-generated platforms to enable students to keep up with the digital advancement of the 21st century and prepare learners for a technologically advanced global landscape [1].

A highly used AI bot in the Moroccan EFL is ChatGPT. It can tailor adaptive personalized learning ways by stocking and analyzing the data of students and, thus, targeting their specific learning needs and styles [12]. Generally, AI bots can foster the EFL learning and teaching experience of Moroccan educators and students in a variety of ways and can thus complement the traditional classroom experience and enhance the quality of the educational content [12]. Bekou et al. [1] explore the opportunities and challenges of implementing ChatGPT in ELT in the Moroccan context. The study namely focuses on three main questions: the perspectives of teachers regarding the use of ChatGPT, its possible benefits for educators and learners alike, and the challenges showcasing related problems. Accordingly, 43.5% of Moroccan educators agree on the usefulness of ChatGPT in ELT (p. 96). The study unveils pivotal benefits related to different areas such as 'personalized learning'. ChatGPT has shown the potential of being a 'teaching aid' through generating engaging authentic conversations for students and providing support for students' weaknesses and instant feedback for their development. Generally, it can help teachers in "designing tests, quizzes, and comprehension passages for classroom practice"(p. 97). Also, they conclude that ChatGPT can help in the enhancement of the professional development of teachers and its continuity by reinforcing their knowledge and memory of pedagogy, curriculum, and teaching procedures. Moreover, ChatGPT can help both educators and students with a resource availability experience with the unlimited access of authentic language materials.

The trend in Morocco is the integration of generative AI into its educational framework. In this transformative scene, the benefits of this bot may turn against its bolstering tenet and synergize with issues. This is in synergy with stressing the ethical guidelines, responsible and correct use [9]. This is a call for educators to maintain accuracy in their implementation of AI in ELT. Pokrivcakova [11] has shown the varied public perception regarding the use of AI in EFL education. The results show that 64.24% support this implementation of AI in university curricula whereas 51.82% consider it a mere threat to students' natural skills. The concerns are namely about AI rendering ELT "less personal" and minimizing the role of teachers. This would be a call for a balanced implementation of AI so that it would complement the human teacher rather than replace them [11]. According to Bekou et al. [1], the concern covers ethical issues such as personal data storing and their future use and security. Ethics also relate to the originality of works for many AI users generate texts and pretend their authorship. Generally, the use of AI bots such as ChatGPT must be aligned with infrastructured awareness on the part of educators to assure their understanding of the potential of AI in ELT [1]. This calls for sufficient guidelines and resources to support AI implementation in the Moroccan ELT domain.

3. Methodology

The present study is an empirical assessment of teachers' use of an AI bot. It is empirical given that the researcher first establishes a hypothesis on the possible problem of AI bot use and uses initially observable data to confirm the hypothesis. It is an assessment given that it puts respondents in a testing condition with certain objectives and procedures seeking their decollement. This study implements both quantitative and qualitative methods of inquiry to unveil the nature and possible falls in Moroccan ELT use of ChatGPT to design formal tests. The present study starts from, first, the de facto recency

of AI integration in ELT and, thus, its recent use by Moroccan teachers, and second, the AI experience that comes with no how-to-use 'catalogue' leaving wide room for impromptu prompts. This paper will examine the validity of Moroccan ELT teachers prompts when asking ChatGPT to design a formal test for the level they are teaching.

3.1. Participants

The data consists of 34 Moroccan ELT teachers working in the private and public sectors in different directorates of the region of Marrakech-Safi. The teachers teach at namely seven levels that follow the education system in Morocco: 1 primary school teacher, 6 first graders teachers, 3 second graders teacher, 6 third graders teachers, 3 common core teachers, 5 first baccalaureate teachers, and 10 second baccalaureate teachers. As for the actual data analysis, it was limited to only 31 respondents given that 3 teachers did not give prompts as requested in the questionnaire.

3.2. Local and Time of the Study

This study has been conducted as an intervention step to guide Moroccan ELT teachers to a proper use of ChatGPT. Teachers belong to different secondary and high schools in the region of Marrakech-Safi. The action continued for three consecutive weeks.

3.3. Instruments and Procedure

To construct this paper, the researcher has first designed a google form to collect data relevant for confirming the hypothesis and building the intervening action. The form had three major questions: 1- What would be the prompt you give to AI (ChatGPT) to generate a formal test for the level you are teaching? 2- What level on the CEFR do you think is the level you are teaching now? 3- Have you respected in your prompt the test construction principles such as aspects of validity, the specification of the tested language level and the test objectives?

Formal assessment includes formal testing administered to a certain level. This assessment should respect specific criteria. Brown [15] specifies such criteria as follows:

"What is the purpose of the test? What are the objectives of the test? How will the test specifications reflect both the purpose and the objectives? How will the test tasks be selected and the separate items arranged? What kind of scoring, grading, and/ or feedback is expected?" [15]

Each of the five questions represents a building block of the administered test. The first question relates to the type of test: placement, diagnostic, summative, etc. When the test designer determines the purpose, they can accustom the objectives accordingly. For instance, once the test designer decides on making a formal test, they should move to choosing the functional expressions, grammatical constructions, language abilities and lexical units to be included. The third question relates to validity in that there must be a proper weighing of tasks. The fourth question relates to practicality, content validity, reliability and authenticity. It must include tasks that reflect what the test-takers have been introduced to. it should be reliably graded by the test corrector. It should contain authentic tasks' content. The final question is about the grading criteria that the test designer must specify.

It is worth noting that the AI bot already creates a formal assessment with the user just typing a simple undetailed prompt without specifying other criteria such as the level or the type of assessment intended (formative, summative, formal, informal, diagnostic, placement, etc.). However, for optimized generated responses, the prompt-writer should follow certain guidelines. The followings are four general guidelines for designing optimized prompts:

1. Give clear and specific instructions
2. Add contextual information and prefixes
3. Adding a partial answer to a prompt
4. Break down prompts into simple components

("Prompt design strategies," n.d.)

The first instruction asks the prompt-writer to tell the AI bot what to do exactly by choosing a prompt that is "clear and specific". Tasks could be summarizing, responding to a question, analyzing, or designing a test in this case. Also, the writer should include certain constraints such as the size of the response and its form. Then, the prompt-writer should specify other requirements of the output such as context specificities. For instance, one should specify the theme, units, level and other details of the prompt. The prompt-designer can include an initial example for the AI bot to start with. Usually, a partial answer begins with a star symbol '*'. The prompt-writer could also try to break down the prompt by first asking the AI bot to do a task as in (1) and then adding information as in (2).

The delivered two questionnaires follow the option of 'multiple pages' with a question per page. The sequence of the questions follows a rational. Once the participant validates an answer, they are transferred to the next page. This was to estop participants from going back to modify their responses after viewing the following questions that would catch the participants' attention to some details that they might have forgotten in their first prompt. Based on the answers of the form, an action was then taken to guide teachers into a more valid and accurate use of ChatGPT to design formal assessment. Participants were provided with a guide showing the necessary details Moroccan ELT teachers should consider when writing a prompt to design a formal assessment. After then, a post-intervention google form was forwarded to participants to track the improvement in their use of the AI bot.

3.4. Data Analysis and Results

Data for this paper were collected at two levels: before the intervention and after the intervention. At the pre-intervention stage, the researcher diagnosed the 'correctness' of the participants' prompts in terms of appropriately detailing the

prompt to spot the possible weaknesses and strengths. This was done by having participants respond to a questionnaire (Google form, Appendix) and testing their proper use. The while-intervention concerned all the participants for a holistic improvement of their use of the AI bot. This was done by providing a guide to be followed that combines both the necessary principles of assessment designing and the basics of prompt designing using the AI bot. As for data collected at the post-intervention stage, the researcher administered a slightly different questionnaire (Appendix) to check the degree of improvement. It should be noted that only 31 respondents have participated with prompts. 3 respondents expressed their disinterest or lack of experience. Thus, the number of analyzed respondents' answers would be limited to 31. The prompt writing must respect two major lines of 'appropriateness': first, creating an effective prompt and second creating a prompt that follows the guidelines of designing a formal assessment. The first line relates to the guidelines of proper use of ChatGPT to have better responses. The second line is related to the test construction principles specifies by Brown [15] in his five questions.

The following table shows the prompts of the 34 participants before the intervention.

Table 1. *The pre-intervention prompts given by 34 participants in this study.*

	Prompts
2 primary school teachers	1. Write a test for primary school students 2. Give a test for first year secondary school students
6 first graders teachers	3. I don't use AI with that concern 4. Give a test for beginners with short exercises. The score is on 10 5. Design a language test for first graders 6. write a test using simple English language 7. generate a simple and short quiz for first graders 8. Give a language test for first graders
3 second graders teacher	9. Write me a text about (topic) using very basic English. 10. Give a test for beginners on how to introduce oneself 11. Write a test for second graders
6 third graders teachers	12. I never used it for that purpose 13. Write a test for 9th grade students on the units of food and drink and house 14. Write a test to assess reading comprehension skills and writing skills… 15. Generate a formal test for ninth grade students on language and writing 16. write me a simple test for beginners 17. write a test for third graders about the unit of Hello
3 common core teachers	18. Write a test for common core student with a short text and comprehension questions on eating habits and a writing section 19. Write a test for common core students on the unit of science and technology 20. Write a test related to the theme of science for common core students
5 first baccalaureate teachers	21. Write a test for first bac students on the units of entertainment and mass media with no writing 22. Generate a test for first bac students 23. Write me a test for first bac students 24. I don't know what to write 25. Give a test for first bac students on the theme of environment
9 second baccalaureate teachers	26. Generate a test for second bac students that covers grammar and writing 27. Generate a test for second bac students on the unit of women and power 28. Write a test for second bac students related to the themes of cultural values and citizenship 29. Write a test with grammar exercises on passive voice and functions exercises on opinion and complaints 30. Write a summative test for 2 bac students that includes writing and reading comprehension

	Prompts
	31. Give a test for second bac students with 2 skills
	32. Write me a language test for second baccalaureate
	33. Generate a test for second bac students
	34. Write a test for second bac students without reading comprehension questions

Table 2. Statement rate of the five basic formal assessment components in the pre-intervention stage.

	The purpose of the test	The objectives of the test	Relevant test specifications	Selection of test tasks and arrangement of test items	The kind of scoring/ grading
Statement rate	6.45 %	12.90 %	100 %	25.80 %	3.22 %
The participants	15 and 30	14, 15, 29, and 30	All participants	4, 14, 15, 18, 26, 29, 30, and 34	4

Table 1 shows the percentage of Moroccan EFL teachers' statement of the five basic constructions of a formal assessment as specified by Brown [15]. The rates were taken at pre-intervention stage. The researcher has tested the assessment-designers ability to combine all five components of a formal test in their prompt. Also, the researcher has examined the prompts to pin point the strategies of prompt-designing the participants have used. It should be noted that 3 participants mentioned that they do not use AI for generating tests. However, they were not excluded from the data gathered at 3 phases given their active participation in the while and post intervention stages. As shown on the table above, most respondents did not clearly state all five components of the test construction. Only 2 participants (6.45 %) mentioned the type of assessment; participant 15 (formative test) and 30 (summative test). As for the testing objectives, only 12.90 % if participants (teachers number 14, 15, 29, and 30) mentioned only a few objectives in namely writing and reading comprehension. Data also show that 31 respondents stated very briefly some specifications. The main recurrent specification was the level of tested students. Whereas other important details such as students' backgrounds or the themes they had studied were mostly out of mention. As for the selection and arrangement of test tasks and items, 8 participants (4, 14, 15, 18, 26, 29, 30, and 34) have mentioned some test sections such as reading comprehension. When it comes to the fifth test construction principle, only one teacher stated it. Participant number 4 has specified the total test score in 10.

It is worth noting that while teachers can, by experience, design a test manually that respects all five components, they may not recognize the fact of not mentioning them when designing a prompt in this line. Some participants' prompts lacked essential components of a test. However, they claimed having included them all.

Based on the data above, the researcher provided participants with a guide targeting the following two elements: designing an effective prompt and using the prompt to design a formal assessment that respects test construction principles. The following table shows port-intervention prompts.

Table 3. The post-intervention prompts given by 31 participants in this study.

	Prompts
2 primary school teachers	1. Write a formal test for A1 students to assess their reading comprehension skills and writing. The score for reading comprehension is 10 and the score for writing is 10 * Reading comprehension 2. Give a formal test for first year secondary school students to test their language knowledge The test should contain three sections: grammar on the verb to be, functions on introducing oneself, and vocabulary on favorite hobbies The score of the test is 10
5 first graders teachers	3. - 4. Give a summative test for beginners with short true or false and multiple choices exercises. The score is on 10

	Prompts
3 second graders teacher	5. Design a formal language test for first graders where the total grade is 20 modify the test to include a grammar section for personal pronouns and verb to be and a writing section on the theme of eating habits where students fill in gaps Adjust the test where the number of questions does not exceed 8 questions. 6. write a test using simple English language for Moroccan students the students are beginners and have just studies the unit of cultural heritage include a 10 lines reading comprehension text with 3 exercises add a vocabulary exercise to assess students understanding of the vocabulary: customs, sculpture, culture shock, ancestors make the total score on 10 7. generate a simple and short quiz for first graders to assess their basic knowledge in numbers, colors, days of the week, age, and nationality. The total quiz score is 10 8. Give a very simple and short reading comprehension quiz for first graders to assess skimming and scanning techniques in a text about school Make the test on 8 points 9. Write me a text about (topic) using very basic English. Provide 4 comprehension exercises for the text. Each of the exercises has 4 pints 10. Give a test for beginners on how to introduce oneself * Reading comprehension
5 third graders teachers	11. Write a diagnostic test for second graders to spot their weaknesses and strengths 12. - 13. Write a test for 9th grade students on the units of food and drink and house Make the test on 20 points and include short and basic exercises 14. Write a diagnostic test to assess reading comprehension skills and writing skills The test is formal and includes 3 exercises for reading comprehension and one paragraph writing section Adjust the test to be on the theme of humor 15. Generate a formal test for ninth grade students on language and writing * fill in the gaps with the right form of the verb to be 16. write me a simple summative test for beginners to assess their knowledge of the simple present and subject an object pronouns 17. write a test for third graders about the unit of Hello include a reading comprehension section with 3 exercises, a grammar exercise on reflexive pronouns and a writing section to introduce themselves the 3 sections must have the grades 6 or 7 or 8 the total score is 20 points
3 common core teachers	18. Write a formal test for common core student with a short text and comprehension questions on eating habits and a writing section The score is on 10 19. Write a formal test for Moroccan common core students on the unit of science and technology to assess reading comprehension skills. The total grade is 10 20. Write a summative test related to the theme of science for Moroccan common core students to assess their knowledge in related vocabulary Include an exercise on present continuous, an exercise on expressing likes and dislikes and a writing section on favorite sciences and scientists The grade is 20 points
4 first baccalaureate teachers	21. Write a test for first bac students on the units of entertainment and mass media with no writing. The grade is 10 * Reading comprehension 22. Generate a norm-referenced test for Moroccan first bac students to assess their comprehension of present perfect and how to express interest or indifference

	Prompts
9 second baccalaureate teachers	23. Write me a summative test for first bac students on the grade of 20 to test their abilities in reading comprehension, grammar, and writing The students are Moroccans who have just studies the units of health and welfare and entertainment 24. - 25. Give a test for first bac students on the theme of environment to assess vocabulary knowledge and the grammatical structures of prepositions of time and place and the past continuous The students are intermediate and Moroccan 26. Generate a short formal quiz for second bac students that covers 2 grammar exercises on modals and writing a book review Adjust the quiz to suit intermediate Moroccan students who have studied the units of women and power and cultural values and issues 27. Generate a 20 points informal test for second bac students on the unit of women and power Include short exercises * Language 28. Write a test for second bac students related to the themes of cultural values and citizenship The test if formal. It includes a section of reading comprehension and a section of writing an article The total grade of the test is 10 29. Write a formal test with grammar exercises on passive voice and functions exercises on opinion and complaints Make the exercises short and easy. The total grade is 20 * Reading comprehension 30. Write a summative test for 2 bac students that includes writing and reading comprehension. The grade is 10 * Reading comprehension 31. Give a test for second bac students with a comprehension text on humor and a writing section on a narrative funny story and a grammar exercise on reported speech 32. Write me a language test for second baccalaureate to assess their understanding of passive voice and writing a report on an event The student's level is between intermediate and upper-intermediate on the CEFR 33. Generate a diagnostic test for second bac students that includes challenging exercises on reading comprehension, grammar, vocabulary, functional expressions, and writing. Modify the test if you know that the students are Moroccan and mostly low achievers Make the total grade on 20 34. Write a formal test for second bac students without reading comprehension questions *Grammar

The table above contains the prompts given by 31 participants in the post-intervention phase. It is necessary to draw attention to the three teachers number 3, 12, and 24 who have not participated in the post-intervention questionnaire though being part of the pre and while-action stages. The table shows the new and optimized prompts given by the same teachers. The following table describes in details the prompts.

Table 4. *Statement rate of the five basic formal assessment components post-intervention.*

	The purpose of the test	**the objectives of the test**	**Relevant test specifications**	**Selection of test tasks and arrangement of test items**	**The kind of scoring/ grading**
Statement rate	64.51 %	77.41 %	100 %	100 %	67.74 %

The table above exhibits the statement rate of the five basic formal assessment components presented by Brown after the

intervention [15]. As shown, all respondents managed to design a prompt that contains most test construction elements if not all. 64.51 % participants expressed very clearly the purpose of the assessment in either 'formal', 'summative', 'diagnostic', or 'norm-referenced' test. Moreover, 77.41 % of the teachers stated the objectives of assessment by specifying a theme or skill. All participants included relevant test specifications in their prompts such as the level of the tested students, their nationality, their knowledge background, weighing of the tasks and test's length. Likewise, 100 % of the respondents made selection of test tasks in sections such as reading comprehension and writing. Also, some of them specified what section to be the introductory in the test such as participants 1, 9, 15, 21, 27, 29, 30, and 34. As for the king of grading, 67.74 % of teachers stated either the total test grade of the distribution of the grade over sections and exercises.

Table 5. *Use rate of the four strategies of prompt-designing during post-intervention.*

	Give clear and specific instructions	Add contextual information and prefixes	Adding a partial answer to a prompt	Break down prompts into simple components
Use rate	100 %	100 %	25.80 %	70.96 %

Table 5 exhibits the use rate of prompt-designing techniques by the 31 participants. As can be seen above, all participants used the first two strategies in that they all gave very clear instructions to ChatGPT on what to do and that they added contextual information to make the generated responses suit more their students. As for the third strategy of adding an initial partial answer to the prompt, only 25.80 % of the respondents made use of it. When it comes to breaking down the prompt to simple components, 70.96 % of the students included subsections to their prompts specifying namely the total test' grade.

4. Discussion

After having administered a pre-action questionnaire (see appendix), the researcher noticed that the respondents exhibit some beginners-performance in writing their test prompts. This is apparent when it comes to first mentioning test principles such as type, purpose, objectives, grading criteria, tasks selection, and the overall practicality, content validity, reliability and authenticity requirements and second writing effective prompts that would lead the AI bot to generate a valid response. Given this performance incompetence, the researcher initiated an action to enable the participants with a concise yet inclusive guide to instruct their prompt-designing methods in constructing a formal assessment. The guide is tripartite. Part one concerns designing optimized prompts (as adapted from "Prompt design strategies," n.d.). This part provides 4 rules to help write effective prompts. Part two includes 5 test construction principles as taken from Brown (pp. 42-43) [15]. The principles target the purpose of the test, its objectives, the test specifications, the test tasks selection and arrangement, and the scoring, grading, and/ or feedback type and criteria. Part three models how to design an effective prompt in terms of general form and specific content. After the action, respondents were subject to the same questionnaire to test the effectiveness of the intervention. Data analysis displayed teachers' enhanced performance in designing their prompts compared to their pre-action performance. On the one hand, they showed aware use of prompt-designing techniques. They have all gave clear instructions and added the needed contextual information to ChatGPT. Also, most of them broke down the prompt into simpler ones to avoid generating unoptimized responses. This exact strategy was never to the-know of any participant as shown in the pre-action data. On the other hand, all respondents managed to design a prompt that contains most test construction elements if not all. Here, and unlike the pre-action results, most respondents mentioned the feedback and grading criteria. Also, all respondents gave detailed test specifications as well as specified the included tasks and their arrangement.

5. Conclusion

The introduction of AI bots is novel to the Moroccan EFL context. Subsequentially, new users would generally lack expertise in dealing with those bots. The present study was conducted to examine the 'correctness' of teachers' in terms of proper prompt-designing and proper use of the prompt to design an assessment that has the necessary features using an AI generative tool, ChatGPT. Assessment is in terms of effective prompt structure and appropriate prompt content. This study is in empirical as it first begins with a hypothesis that AI bot users in the Moroccan EFL context lack effective how-to-use strategies possible problem and uses pre-action attested data to confirm the hypothesis. This study is an assessment because it 'tests' respondents' performance. The testing has specific objectives and procedures to enhance respondents' ChatGPT use. The study is a quantitative and qualitative one given that data are measured in terms of rates and are described in terms of content.

In general, the pre- and post-action results discrepancy con-

firm the assumption that Moroccan EFL teachers prompt-designing strategies need assistance to improve on them. It shows that though AI bots seek assistance of users in performing their tasks, test construction in this case, however, their use is taken for granted as users ignore existing guidelines showing maximized effectiveness in use. This presumed knowledge of use is what gives unoptimized generated responses by the AI bot. The present research sheds light on this reality and seeks to improve on the users' performance.

6. Practical Implications

It can be concluded from this study that Moroccan EFL teachers show some performance deficiency in using ChatGPT in designing formal tests. This research provides a practical guide to enhance their performance. It is recommended that teachers be trained in how to use AI bots for teaching and planning purposes. Moroccan EFL teachers are trained in ICT use but, given the novelty of AI integration, it has not yet been included in their professional training. Another implication is that educational policy makers can monitor the use of AI by teachers to improve on their teaching outcomes.

Abbreviations

AI	Artificial Intelligence
EFL	English as a Foreign Language
ELT	English Language Teaching

Conflicts of Interest

The authors declare no conflict of interest.

Appendix

Pre-Intervention Questionnaire

An Empirical Assessment of Moroccan EFL Teachers' Use of Generative AI for EFL Formal Assessment

This google form is designed as part of an empirical investigation into the use of an AI tool (ChatGPT) by Moroccan EFL teachers in generating formal tests to measure the language knowledge of their students. The form is composed of no more than 4 main questions. Teachers are kindly required to be direct, brief, and honest in their replies for adequate objectivity. Teachers are also asked to provide their emails for a possible contact during the study.

Email Address

What level are you teaching?

1. Pre-school
2. Primary school
3. First Graders
4. Second Graders
5. Third Graders
6. Common Core
7. First Year Baccalaureate

Second Year Baccalaureate

What would be the prompt you give to AI (ChatGPT) to generate a formal test for the level you are teaching?

What level on the CEFR are you teaching?

A1 Elementary level

A2 Pre-intermediate

B1 Intermediate level

B2 Upper intermediate level

C1 Advanced level

C2 Proficiency level

Have you respected in your prompt the test construction principles such as aspects of validity, the specification of the tested language level and the test objectives?

The guide (While-intervention):

Part one: designing an effective prompt

For optimized generated responses, the prompt-writer should follow certain guidelines. The followings are four general guidelines for designing optimized prompts:

Give clear and specific instructions

Add contextual information and prefixes

Adding a partial answer to a prompt

Break down prompts into simple components

("Prompt design strategies," n.d.)

The first instruction asks the prompt-writer to tell the AI bot what to do exactly by choosing a prompt that is "clear and specific". Tasks could be summarizing, responding to a question, analyzing, or designing a test in this case. Also, the writer should include certain constraints such as the size of the response and its form. Then, the prompt-writer should specify other requirements of the output such as context specificities. For instance, one should specify the theme, units, level and other details of the prompt. The prompt-designer can include an initial example for the AI bot to start with. Usually, a partial answer begins with a star symbol '*'. The prompt-writer could also try to break down the prompt by first asking the AI bot to do a task as in (1) and then adding information as in (2).

Part two: test construction principles

Formal assessment includes formal testing administered to a certain level. This assessment should respect specific criteria. Brown specifies such criteria as follows [15]:

"What is the purpose of the test? What are the objectives of the test? How will the test specifications reflect both the purpose and the objectives? How will the test tasks be selected and the separate items arranged? What kind of scoring, grading, and/ or feedback is expected?" (Brown, pp. 42-43) [15]

Each of the five questions represents a building block of the administered test. The first question relates to the type of test: placement, diagnostic, summative, etc. When the test designer determines the purpose, they can accustom the objectives accordingly. For instance, once the test designer decides on making a formal test, they should move to choosing the functional expressions, grammatical constructions, language abilities and lexical units to be included. The third question

relates to validity in that there must be a proper weighing of tasks. The fourth question relates to practicality, content validity, reliability and authenticity. It must include tasks that reflect what the test-takers have been introduced to. it should be reliably graded by the test corrector. It should contain authentic tasks' content. The final question is about the grading criteria that the test designer must specify.

Part three: an example of how to use the prompt to design a formal assessment that respects test construction principles.

Prompt: Design a summative formal assessment for 2 bac students to test grammatical constructions, functional expressions, vocabulary, reading comprehension and writing

* Reading comprehension

Prompt: Include a 10 lines reading comprehension text with 5 comprehension questions, one exercise with 3 questions for the grammar section, one gap filling-in exercise for functions, etc.

Prompt: Adjust the assessment to suit Moroccan 2 bac students whose level in intermediate and who have studied the units: A, B, and C.

Prompt: Adjust the assessment where grammar, vocabulary, and functions are in one section called Language

Prompt: Include grades for each question where the total score of the whole test 20

Prompt: Adjust the grades where each of the three sections has the grade: 6 or 7 or 8.

Here, all five test construction principles have been respected (content of each section such as functions has not been mentioned. This does not mean that the prompt-writer should ignore them). Besides, all four strategies of prompt-designing have been used. The prompts relied mainly on the fourth strategy to guarantee generating relevant and valid responses.

Post-intervention Questionnaire

An Empirical Assessment of Moroccan EFL Teachers' Use of Generative AI for EFL Formal Assessment

This google form is designed as part of an empirical investigation into the use of an AI tool (ChatGPT) by Moroccan EFL teachers in generating formal tests to measure the language knowledge of their students. The form is composed of no more than 2 main questions. Teachers are kindly required to be direct, brief, and honest in their replies for adequate objectivity.

Email Address

What was the prompt you gave to AI (ChatGPT) to generate a formal test for the level you are teaching before the guide?

What would be the prompt you give to AI (ChatGPT) to generate a formal test for the level you are teaching now?

References

[1] Bekou, A., Ben Mhamed, M., & Assissou, K. (2024). *Exploring opportunities and challenges of using ChatGPT in English language teaching (ELT) in Morocco*. Focus on ELT Journal, 6(1), 87–106. https://doi.org/10.14744/felt.6.1.7

[2] Buchanan, Bruce. (2005). A (Very) *Brief History of Artificial Intelligence*. AI Magazine. 26. 53-60.

[3] Gregersen, E. (2023). *ChatGPT*. Encyclopedia Britannica. Retrieved March 26, 2024, from https://www.britannica.com/technology/ChatGPT

[4] Halaweh, M. (2023). *ChatGPT in education: Strategies for responsible implementation*. Contemporary Educational Technology, 15(2), ep421. https://doi.org/10.30935/cedtech/13036

[5] Hughes, A. (2023). *ChatGPT: Everything you need to know about Open Ai's GPT-4 tool.* BBC Science Focus Magazine - science, nature, technology, Q&As. https://www.sciencefocus.com/future-technology/gpt-3

[6] Monostori, L. (2014). *Artificial Intelligence*. In: Laperrière, L., Reinhart, G. (eds) CIRP Encyclopedia of Production Engineering. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-20617-7_16703

[7] McCarthy, J. (2007). *What Is Artificial Intelligence?* Technical report, Stanford University, Available online at: http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html (Accessed June 2, 2018).

[8] Copeland, B. (2024). *Artificial Intelligence*. Encyclopedia Britannica. https://www.britannica.com/technology/artificial-intelligence

[9] *AI and education: "Morocco is not the country to miss out".* Interview with Dr Abdelilah Kadili. (2024, March 29). Human Technology Foundation. https://www.human-technology-foundation.org/news/ai-and-education-morocco-is-not-the-country-to-miss-out-interview-with-dr-abdelilah-kadili

[10] Bigaj-Kisala, M. (2023). *AI in the classroom?* Interview with ChatGPT. That is Evil. https://thatisevil.education/2023/02/ai-in-the-classroom-interview-with-chatgpt/

[11] Pokrivcakova, S. (2023). *Pre-service teachers' attitudes towards artificial intelligence and its integration into EFL teaching and learning*. Journal of Language and Cultural Education, 11(3), 100-114. https://doi.org/10.2478/jolace-2023-0031

[12] Qbaibi, S. (2023). *Unlocking the future of education in Morocco: How AI could revolutionize learning*. Retrieved April 4, 2024, from https://www.moroccoworldnews.com/2023/07/356679/unlocking-the-future-of-education-in-morocco-how-ai-could-revolutionize-learning

[13] Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests.* Oxford University Press.

[14] Black, P., & Wiliam, D. (1998). *Inside the Black Box: Raising Standards Through Classroom Assessment.* Phi Delta Kappan, 80(2), 139-148.

[15] Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices*. Pearson Education.

[16] Gronlund, N. E., & Waugh, C. K. (2009). *Assessment of Student Achievement*. Pearson.

[17] Hebert, T. (2001). *Portfolio Assessment and the National Standards*. English Teaching Forum, 39(4), 18-27.

[18] Messick, S. (1989). *Validity*. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13-103). Macmillan.

[19] Richards, J. C., & Schmidt, R. (2002). *Longman Dictionary of Language Teaching and Applied Linguistics* (3rd ed.). Pearson Education.

[20] Harmer, J. (2015). *The practice of English language teaching*. Pearson Longman.