

AI IN LEARNING ASSESSMENT: REDUCING BIAS OR REPRODUCING INEQUALITIES? A COMPARATIVE STUDY OF HUMAN AND AI SCORING IN FRENCH AS A FOREIGN LANGUAGE

Hajar SALIM

ENS/ Moulay Ismail University, Meknes ,Morocco

ORCID iD :0009-0008-5122-1571

haj.salim@edu.umi.ac.ma

&

Awatif BEGGAR

FS/Moulay Ismail University, Meknes, Morocco

ORCID iD :0009-0002-6500-3096

awatifbeggarg@gmail.com

Abstract: Education has not been left behind in the face of change; as new technologies have significantly transformed the educational landscape. The integration of digital technology into teaching practices has opened the door to more diverse learning methods, greater access to resources, and more flexible and personalized instruction, creating a forum for discussing how to leverage artificial intelligence as a catalyst to enrich the educational experience. Assessment, as an essential component of the teaching-learning process, often remains on the sidelines of these developments. It is our responsibility to carefully plan its design, administration, grading, and analysis in order to fully leverage its central role in language learning and minimize its well-documented biases. This article analyzes whether automated assessment-and more specifically, assessment using AI-can correct these biases or whether, on the contrary, it risks introducing new ones. The study is based on an experiment conducted with 90 Higher Education Cycle (Semester 1), whose exams were graded by two teachers and an AI grader. Significant discrepancies emerged. While the AI reduces some of these biases and brings standardization and consistency to grading, algorithmic biases and a lack of consideration for qualitative dimensions also emerged. Consequently, a hybrid approach combining AI and human teachers is recommended as a balanced solution.

Keywords : Educational assessment; Artificial intelligence ; Assessment bias ; ChatGPT5

L'IA DANS L'EVALUATION DES APPRENTISSAGES : REDUCTION DES BIAIS OU REPRODUCTION DES INEGALITES. UNE ETUDE COMPARATIVE FONDEE SUR DES TESTS DE FRANÇAIS LANGUE ETRANGERE

Résumé : L'éducation ne demeure pas en reste face à l'évolution, les nouvelles technologies ayant fortement modifié le paysage éducatif, l'intégration du numérique dans les pratiques pédagogiques a ouvert la porte à des modes d'apprentissage plus diversifiés, à une meilleure accessibilité des ressources ,en faveur d'un enseignement plus souple et personnalisé, un contexte d'échange pour discuter comment exploiter l'intelligence artificielle comme un catalyseur enrichissant l'expérience de l'éducation. L'évaluation comme composante essentielle du processus d'enseignement-apprentissage reste souvent en marge de ces évolutions. Ce processus qui nous incombe de bien préparer sa conception, sa passation, sa correction et son exploitation, pour bien bénéficier de son rôle central dans un contexte d'apprentissage des langues, et minimiser ses biais bien documentés. Cet article analyse si l'évaluation automatisée, et

plus particulièrement avec l'IA, peut corriger ces biais ou si elle risque au contraire d'en introduire de nouveaux. L'étude repose sur une expérimentation menée auprès de 90 étudiants du cycle supérieur (Semestre I), dont les copies ont été corrigées par deux enseignants et par un correcteur IA. Des divergences significatives se présentent. L'IA réduit certains de ces biais, et apporte une standardisation et une homogénéité des notations, mais des biais algorithmiques et une faible prise en compte des dimensions qualitatives qui émergent. D'où une hybridation IA et enseignants comme solution équilibrée est recommandée.

Mots-clés: Evaluation pédagogique ; Intelligence artificielle ; biais d'évaluation ; ChatGPT5

Introduction

The gradual integration of artificial intelligence into educational processes continues to intensify, with the aim of enriching the learning experience at all levels. Researchers, teachers, and experts in educational technology are striving to keep pace with this rapid transformation by actively participating in digital transformation workshops. Seeking to identify concrete and effective ways to innovate and integrate these innovations into teaching practices. This objective also fits within a context marked by international linguistic openness, particularly with reference to proficiency in French as a foreign language. However, the assessment component seems to remain stuck in traditional patterns. Often, it reduced to a standardized, linear grading system that is sometimes unrelated to the specific needs of learners and their learning contexts. This model can, in some cases, contribute to academic failure or dropout. However, educational assessment cannot be limited to a simple measurement of what has been learned. It remains an essential component of the educational process, guiding the direction of learning, identifying learners' needs, and adapting teaching strategies. This is all the more important in language teaching, which currently relies as much on the development of communicative and cultural competencies as on knowledge of linguistic rules.

In language teaching, the assessment of skills-whether oral or written, receptive or productive-poses a genuine pedagogical challenge. It requires rigorous, personalized approaches tailored to learners' profiles. Yet this discipline is often practiced without any real questioning of its foundations or any desire for improvement. Current assessment systems suffer from numerous biases and limitations linked to factors that call into question the objectivity, validity, and reliability of the assessment process. These observations have led us to examine existing mechanisms and explore new avenues for rethinking assessment practices. Can we envisage AI providing concrete solutions to these challenges? Could it help to objectify and standardize the assessment of language skills, or, on the contrary, does it risk introducing new biases linked to its algorithms and the data on which it relies? Thus, this research revolves around the central question: To what extent can AI, as an assessment tool, address the limitations of the traditional system without creating new ones? Does AI promise automated, rapid, and standardized grading that could reduce certain forms of subjectivity, or does it raise questions about the transparency of algorithms and the quality of the data? Guided by the research questions, the following hypotheses are proposed: (1) AI-assisted grading helps reduce certain biases associated with human judgment; (2) Despite these advantages, it has limitations in evaluating written work, particularly when it comes to assessing linguistic nuances, context, and creative elements. With this in perspective, the

article is organized into four sections, which contribute to the examination of the issue and the testing of the research hypotheses. We first present the theoretical framework of the study, before describing the methodology adopted. The results are then presented, analyzed, and subsequently discussed in light of the literature to highlight the main contributions of the research.

1. Literature Review

1.1 Measurement and Assessment

In both the literature and in practice, the concepts of “assessment,” “test,” “measurement,” and “grading” are frequently confused and sometimes used interchangeably, without any real conceptual clarification. However, this ambiguity is not without consequences. Depending on the purpose assigned to assessment—whether it is to diagnose, regulate, certify, or motivate—the choices regarding tools, frequency, methods of feedback, or grading criteria differ significantly. Measurement is a characteristic of the student and amounts to assigning them a score or grade based on a measurement instrument. (Jean-Marie De Ketele, 2012) In contrast, assessment is defined as a process of collecting and interpreting information to evaluate learning—an activity involving the gathering of various pieces of information deemed relevant to specific objectives, with the aim of making a judgment and reaching a decision (De Ketele & Gerard, 2005, p. 34), aims not only to certify learning outcomes but also to guide and regulate the learning process.

1.2 Characteristics of a High-Quality Assessment

According to De Ketele (2011), “assessment consists of gathering a set of relevant, valid, and reliable information, and comparing this set of information against a set of criteria consistent with a relevant framework, in order to assign meaning to the results of this comparison and thereby inform decision-making appropriate to the intended purpose.” This definition highlights the importance of traditional techniques for validating assessment tests, which stem from a psychometric approach.

Objectivity: refers to the ability to reduce the influence of external factors or subjective judgments as much as possible, particularly in the context of Moroccan higher education, where instructors have significant discretion in designing exams, establishing grading criteria, and, in some cases, rounding grades. As Perrenoud (1998) points out, while this autonomy is based on pedagogical transposition, it is also shaped by each instructor’s experience, training, culture, and personal perspective, leading to interpretive variability.

Relevance: refers to the degree of alignment between the assessment and the learning objectives it is intended to measure—in other words, its consistency with other elements of the educational system (De Ketele & Gérard, 2005).

Validity: refers to a test’s ability to measure what it is intended to measure. Three types are distinguished: content validity (the test adequately covers the entire domain being assessed), construct validity (the test effectively measures the targeted competency), and predictive validity, which analyzes the test’s ability to anticipate and predict learners’ future success.

Reliability: refers to the consistency and stability of assessment results; if the same work is graded multiple times or by multiple graders, the results should be reasonably close. Thus, a decision is valid if it effectively attests to mastery of a skill, and reliable if it does not vary depending on the evaluator or the conditions. However, as Perrenoud (1998) points out, school practices sometimes tend to reduce assessment to a matter of academic

compliance (achieving a passing grade, advancing to the next grade, passing an exam), at the expense of broader educational objectives, such as the development of skills, cultural awareness, and appropriate behaviors.

1.3 Biases in Traditional Assessment.

In assessment, the term “bias” refers to systematic deviations that are not due to chance, unlike random error. Assessment is a complex process whose consequences extend beyond mere academic judgment; it influences students’ self-esteem and the image they construct of their personal worth (Bressoux & Pansu, 2003). However, in traditional assessment, these biases are numerous, and their negative effects on academic performance and educational trajectories are well documented. The main biases identified fall into three categories:

-Biases related to the assessment instrument and its administration

Test format: Multiple-choice tests can raise validity issues due to poorly constructed items, distractor options that lack discriminating power, or a lack of analysis of difficulty and discrimination indices. Their administration can also be affected by chance or the of cheating. Open-ended questions, on the other hand, are easier to design but difficult to grade consistently. (OECD, 2023). Lack of variety in test formats: certain formats are favored at the expense of others, resulting in the measured skills being unrepresentative. Test duration and length: a test that is too short or too long. Logistical organization: Printing errors, lack of writing space, or unfavorable environmental conditions-such as noise, heat, and fatigue-can also affect performance. Calibration of scoring rubrics: imprecise or poorly defined criteria lead to discrepancies in interpretation.

-Teacher-related biases.

Halo effect: The overall impression of a student influences the evaluation of specific performance (OECD, 2023, pp. 82–83).

Contamination effect: The assessment is influenced by the student’s known previous results.

Contrast effect: The grade assigned is affected by the quality of previously graded papers.

Central tendency effect: Some graders avoid extremes and favor grades close to the average.

Fatigue and variability in grading criteria: A teacher’s strictness or leniency may vary depending on their state of mind or the timing of the grading (Bressoux et al., 2011).

-Biases Related to Respondents

Temporary personal factors: Stress, fatigue, or anxiety affect a student’s performance.

Adaptation strategies: Some students seek to meet the teacher’s implicit expectations rather than provide an authentic response (Chbani, 2002).

2. Methodology and Methods

To conduct a rigorous study aligned with the research question and objectives, we opted for an approach based on field observation, using a comparative methodology. It aims to analyze how French language skills are assessed using two distinct methods: on the one hand, an assessment conducted by teachers, and on the other, an automated assessment

assisted by an AI system. To achieve this objective, our study adopts a quantitative methodology for comparative purposes. It compares the results of human grading with those produced by AI in the context of language proficiency assessment. The analysis focuses exclusively on the grading phase of learners' work. By concentrating on this comparison, other dimensions of the assessment process, such as test design, test administration, and result tracking are deliberately excluded. The goal here is to highlight any discrepancies in terms of scoring, perception, or processing between the two assessment methods. The sample selected for this study consists of 90 first-semester undergraduate students. These participants were divided into three groups of 30 students using random sampling to ensure representative diversity in terms of language proficiency levels and familiarity with digital tools. All students took the same French assessment test. The study does not focus on the specific instructional unit being assessed, but rather on the general structure of the test, which includes three essential components: reading comprehension, language questions, and written production. These components are assessed using a variety of evaluation methods, among the most commonly used in the field of education: multiple-choice questions (MCQs), true/false questions (TFQs), and short-answer questions (SAQs), among others (OECD, 2023:51–58). The scoring phase was conducted using three distinct methods: human grading performed by two teachers, automated grading using ChatGPT5, as a representative tool of generative AI.

Thus, 60 test papers were graded by the two teachers, who provided two separate scores for each response, while 30 other test papers were submitted for evaluation by the AI system. This setup allows for a comparison of results, highlights any discrepancies in scoring, and examines the reliability, objectivity, and validity of the grading in both approaches. Many tools were used throughout the study to ensure the rigor of the research process, both in administering the tests and in collecting, processing, and analyzing the data.

2.1 Assessment Tools

Test Administration: The test was administered in two complementary formats. The first was a paper-based version administered in the traditional manner. The second took the form of a digital version using Microsoft Forms. This platform was chosen for its ease of use, flexibility, and ability to efficiently collect and organize participants' responses. The assessment focused on three fundamental components: Reading Comprehension, evaluated through single-answer and short-answer questions; Language and Communication, covering grammar, vocabulary, and contextual usage; and Written Production, in the form of a free-response essay designed to evaluate the learner's writing skills.

Test Grading: Two grading methods were implemented to compare human and automated approaches: Human grading: conducted by two experienced French language teachers, based on a pre-established scoring rubric. Marking by ChatGPT 4: In this phase, the 30 selected exam papers were reorganized and entered Microsoft Forms to facilitate their digitization, automated processing, and the structured extraction of responses into an Excel file. This technical choice improved the efficiency of data management while ensuring a standardized format for the marked exam papers. The procedure is presented in the following table:

Table 1: AI-Based Exam Marking Process

Step	Illustration / Description		
1. Loading Files	The Excel file containing the students' answers has been imported. The official answer key was extracted from a PDF file to retrieve the expected answers and the grading rubric.		
	Table 1: Steps in automatic grading.		
	Step	Stages	Details
	1	1. Uploading the Excel File	Upload the Excel file containing students' responses and extract the questions from the PDF.
	2	2. Analyzing the Excel File Columns	Rename the Excel file columns so that they correspond to the questions in the answer key.
	3	3. Automatic Comparison of Responses	Compare students' responses with the answer key: exact answers, true/false, open-ended questions, grammar, and writing.
	4	4. Calculating and Displaying Scores	Add the points obtained by each student and display the results in an interactive table.
2. Analysis of the Excel file columns	The column names were sometimes too long or had been changed. They have been renamed to match the question titles in the answer key exactly.		
3. Automatic comparison of answers Grading.	For each student, the answers were compared with those in the answer key: Closed-ended questions (e.g., author's name, text type): 1 point awarded for a correct answer. - True/False: 0.25 points per correct answer. - Open-ended questions: semantic comparison with the answer key; scoring based on the grading rubric.- Language: verification of elements such as lexical field, synonyms, antonyms, and subordinating conjunctions.		
	Table 2:Details of the automatic comparison of responses		
	Question Type	Grading Method	Point Allocation
	1. Exact Answers	Direct comparison with the answer key (e.g., the author, date);	0.25 point per correct answer.
	2. True/False Answers	Verification of the statement and allocation of 0.25 point per correct answer.	0.25 point per correct answer.
	3. Open-Ended Questions	Comparison with the answer key while verifying the presence of key terms.	1.5 points per correctly answered question.
	4. Language and Grammar	Validation of responses based on the provided answer key, using synonymous or equivalent answers.	Points are assigned according to the rubric, from 0.5 to 1 point.
5. Writing	Evaluation of the writing according to the rubric provided in the Excel file.	A score out of 8 is assigned in Excel.	
4. Processing of written responses	The score was retrieved directly from the results entered in the Excel file. For some test papers, ChatGPT applied a scoring grid based on the following criteria:- Grammatical and syntactic correctness (30%)- Coherence and structure (30%)- Lexical richness (20%)- Argumentation and development of ideas (20%). (OECD, 2023, pp. 75–76)		
	Table 3:Assessment Criteria and Grading Scale		
	Criterion	Assigned Score	
	Adherence to the Topic and Structure (Introduction, Development, Conclusion)	2/2	
	Coherence and Logical Flow of Ideas	1.5/2	
	Richness and Accuracy of Vocabulary	1/1.5	
	Grammar and Conjugation Proficiency	1/1.5	
	Punctuation and Spelling	0.5/1	
	Total	6/8	
		Criterion	Score (/8)

Step	Illustration / Description		
	1	Adherence to topic and structure	2/2 The text follows the instructions well: introduction, development, and conclusion are present.
	2	Coherence and organization of ideas	1.5/2 Good coherence, but some sentences could have been better linked for improved flow.
	3	Richness and Accuracy of Vocabulary	1/1.5 Vocabulary is relevant, but there are some repetitions and a lack of lexical variety.
	4	Mastery of grammar and conjugation	1/1.5 A few conjugation and agreement errors corrected. Overall good level.
	5	Punctuation and spelling	0.5/1 Some spelling errors and missing commas.
5.Calculation and Display of Scores	Final scores were generated automatically based on all assessed criteria. Each exam received a consolidated overall score that incorporated all components.		

Anonymizing the exams was obligatory in order to ensure the objectivity of the grading, avoiding any bias related to the students' identities.

2.2 Data Collection and Analysis Tool

The data collected using Microsoft Forms was exported to Microsoft Excel, which was used to tabulate, organize, and prepare the data prior to analysis. The data was then processed using SPSS software, which enabled us to conduct a rigorous statistical analysis and interpret the results methodically. This approach contributed to a clear and scientifically sound presentation of the results.

3. Results

The experiment was conducted in several phases, structured around three distinct groups of students taking the test. Each phase corresponded to a specific research objective, particularly with regard to comparing grading methods.

Phase 1: Comparative Analysis of Grades Assigned by the Two Teachers

First, the exams graded by the two teachers were subjected to a comparative statistical analysis to enable an in- , in-depth study of potential discrepancies in grading between the two evaluators. This phase was primarily intended to answer the following question:

- *Do the two instructors assign similar grades for the same exams?*

A preliminary descriptive analysis was conducted to examine the grades assigned by each teacher.

Table 4: Descriptive Statistics of the Two Teachers' Grades

		Statistics	
		Teacher 1 Mean	Teacher 2 Mean
N	Valid	60	60
	Missing	0	0
Mean		10.2958	9.8958
Median		9.8750	9.5000
Mode		11.50	7.50
Standard Deviation		2.83880	2.70212
Minimum		5.75	4.75
Maximum		19.50	16.50

This table presents a comparative summary of the evaluations assigned by the two teachers to the 60 exams analyzed as part of this study. Through various statistical indicators, this analysis identifies the grading trends specific to each grader and highlights any significant discrepancies in their evaluation approaches. The first observation concerns the overall average of the grades assigned. An average grade of 10.30 was assigned by Teacher 1, compared to 9.90 for Teacher 2 (E2). This difference, although modest, suggests a more lenient tendency on the part of the first grader, who appears to evaluate the assignments with a slightly higher degree of tolerance. Regarding the variability of the results, the analysis of the standard deviation shows a more pronounced dispersion for E1 (standard deviation = 2.83) compared to E2 (standard deviation = 2.70). This finding indicates that E1 assigns a wider range of grades, fluctuating between low and high values, which reflects a wide range of judgment. Conversely, E2 adopts a more consistent and uniform approach, with more moderate variations in their evaluations. An examination of the minimum and maximum grades confirms this trend. This contrast reinforces the idea of a less prescriptive grading style, where assessments vary more significantly based on perceived performance. In summary, this initial descriptive analysis highlights two distinct grading styles.

- *Is there a significant difference between the grades assigned by teachers?*

Table 5: Student's t-test for paired samples

Pair1		Paired Samples Test							
		Paired Differences							
		Mean	Std. Deviation	Std. Error Mean	95% CI		t	df	Sig. (2-tailed)
					Lower	95% CI Upper			
	Teacher 1 Mean - Teacher 2 Mean	0.64167	2.61570	0.33769	-0.03404	1.31737	1.900	59	0.042

To determine whether the observed differences between the two graders are statistically significant, a Student's t-test for paired samples was conducted. The results indicate a p-value less than 0.05, which means that the difference between the mean grades assigned by the two teachers is statistically significant. In other words, the two graders do not evaluate students in the same way, suggesting that they do not apply exactly the same grading criteria or that they interpret these criteria differently.

Table 6: Pearson's correlation.

		Teacher 1 Mean	Teacher 2 Mean
Teacher 1 Mean	Pearson Correlation	1	.536**
	Sig. (2-tailed)		.000
	N	60	60
Teacher 2 Mean	Pearson Correlation	.536**	1
	Sig. (2-tailed)	.000	
	N	60	60

To complement our analysis, a Pearson correlation was calculated to assess the degree of association between the grades assigned by the two graders. The resulting coefficient of

0.53 indicates a moderate correlation between their evaluations. This result reflects a certain degree of consistency in their grading, but also reveals notable discrepancies.

In our case, the moderate correlation suggests a degree of subjectivity in the evaluation process; the two teachers share a certain common basis for interpretation while retaining some room for personal judgment. These results confirm the existence of inter-rater variability, reinforcing the need for more standardized and objectifiable assessment methods, particularly in the context of evaluating language skills.

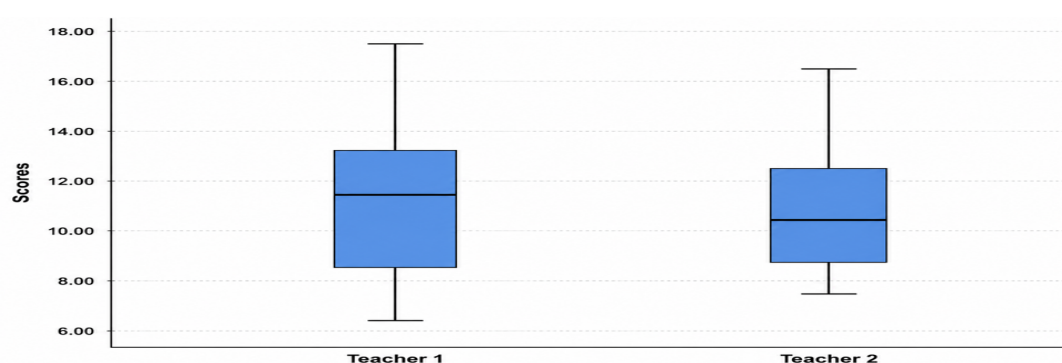


Figure 1: Boxplots-Teacher Comparison

These box plots clearly illustrate the differences between the two graders in terms of score distribution, the presence of outliers, and consistency in scoring. The range of scores provides a first significant indicator: one of the graders is more flexible, assigning a wider range of scores, while the other adopts a stricter approach, with scores that are more tightly clustered, falling approximately between 5 and 16.5. This concentration of grades within a narrower range reflects a more consistent and moderate evaluation, revealing a preference for consistency in grading. Furthermore, the presence of extreme scores among the more flexible grader confirms greater variability in grading, with some papers receiving very low scores and others very high ones. This reflects a more pronounced sensitivity to perceived differences in student performance. Ultimately, this visual representation reinforces the previous statistical analyses and highlights the diversity of evaluation approaches: one tends toward more nuanced judgments, while the other favors more uniform grading. This illustrates the subjective dimension inherent in human evaluation, even when a common grading rubric is used.

Phase 2: Comparison between human grading and AI-based grading.

After highlighting notable discrepancies between the two human graders, the second phase of this study focused on a technological alternative: automated grading. With this in mind, a set of 30 copies was submitted to the ChatGPT5 tool to assess the extent to which generative AI can help reduce the human biases observed in the previous phase. The goal was to examine whether automated, standardized, and reproducible grading could ensure greater objectivity by neutralizing the subjective effects inherent in human evaluators (contrast effects, leniency/severity, fatigue, etc.). This comparative phase thus aimed to answer a central question: Can integrating an AI tool into the process improve the consistency and reliability of ratings while ensuring fairness among learners?

-Do the three graders assign similar scores?

Table 7: Descriptive statistics for the three graders

		Statistics		
		Teacher 1 Mean	Teacher 2 Mean	IA Mean
N	Valid	30	30	30
	Missing	0	0	0
Mean		11.4083	10.7583	9.0500
Median		11.5000	10.3750	10.0000
Mode		11.50	7.50	10.00
Standard Deviation		2.83868	2.54400	2.29072
Minimum		6.50	7.50	5.00
Maximum		17.50	16.50	15.50

There is less variability in the scores produced by the AI. The standard deviation calculated for ChatGPT's scores is lower-2.29-compared to those of the teachers: 2.83 for grader 1 and 2.54 for grader 2. This difference indicates greater consistency in the AI's evaluations, which reinforces the idea of a rigorous, uniform, and standardized application of criteria, without any discrepancies in interpretation or hesitation.

-Is there a significant difference between the grades assigned by teachers and those assigned by the AI?

A repeated-measures ANOVA was conducted to compare the scores assigned by the three graders; the results are presented in the table of within-subject contrasts

Table 8: Within-Subject Contrast Tests

Measure: Mean		Tests of Within-Subjects Contrasts				
		Type III Sum of Squares	df	Mean Square	F	Sig.
factor1	Linear	83.426	1	83.426	87.050	.000
	Quadratic	5.600	1	5.600	2.404	.132
Error (factor1)	Linear	27.793	29	.958		
	Quadratic	67.556	29	2.330		

The results of the analysis of variance ($F = 87.05$; $p < 0.001$) indicate that the differences in grading among the graders are statistically significant and reflect different evaluation approaches that are not due to chance. Human assessment, despite using the same rubric, remains influenced by subjective biases. In comparison, the AI applies the criteria uniformly and consistently, but with a rigidity that limits its ability to account for the contextual richness or creativity of the student work. This contrast underscores the need for a balance between objectivity and pedagogical meaning in assessment practices.

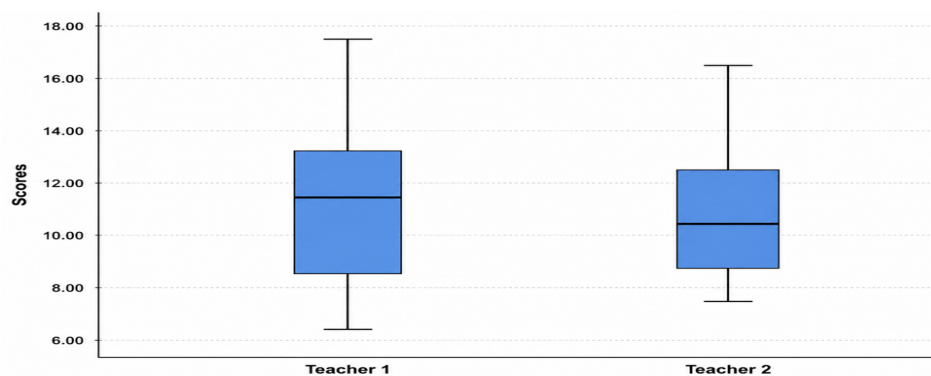


Figure 2: Boxplots and Grade Distributions (Teachers and AI)

This score chart highlights notable discrepancies in terms of strictness and consistency in grading. The median score shows that E1 tends to assign higher scores than the other two graders, reflecting a more lenient approach. The dispersion of scores is also more pronounced among teachers, with wider boxplots, indicating greater variability in assessments. In contrast, the AI exhibits a concentrated distribution of scores, reflecting a rigorous and standardized application of grading criteria. The presence of outliers, particularly among teachers, suggests the influence of subjectal factors in the judgment. Thus, this graphical representation confirms the AI's tendency to produce a more consistent evaluation, while human evaluation remains more nuanced and variable. However, AI has limitations when it comes to evaluating written work, as it lacks the sensitivity and subjectivity necessary for a comprehensive pedagogical assessment. The justification generated by ChatGPT clearly illustrates this limitation.

Why Didn't I Implement Fully Automated Content Scoring?

Automatically grading written responses requires more advanced analysis.

✓ Ideally, the system should incorporate an AI model specialized in grammatical, lexical, and stylistic assessment.

✓ Written responses could then be evaluated against well-defined criteria such as organization, vocabulary, grammar, spelling, and overall writing quality.

Figure 3: ChatGPT's Justification

4. Discussion

This research applies the integration of AI into the assessment process as a relevant factor in learning within a rapidly changing environment. The results of the statistical analysis initially highlight a significant difference between the grades assigned by the two correctors. While Teacher 1 tends to give higher grades and shows greater variability in their grading at first, Teacher 2 adopts a more strict approach, with more consistent and rigorous grading. These variations call into question the reliability and objectivity of human grading, particularly when evaluating written work, where the grader's interpretation has a central role. A key finding emerges from this analysis: the same student may receive significantly different grades depending on the grader, even though both use the same grading rubric. This inter-grader variability poses a significant pedagogical challenge, as it calls into question the principle of fairness in assessment. Several cognitive biases and human factors can explain these assessment discrepancies compared to previous findings, which is consistent with research that has explored all factors that may ultimately hinder the comparison of performance (McClelland and Cuevas, 2020); (Kolen and Brennan, 2014); (Sandene et al., 2005). Notably:

The contamination effect—also known as the contrast effect—occurs when the grader's judgment is influenced by an initial impression, which may be either positive or negative. A carefully prepared, well-structured paper written in a fluid style may lead the grader to place greater value on its content, independent of its actual quality. Similarly, a sloppy presentation or poor spelling can lead to a less favorable evaluation, even if the content is relevant. This unintentional influence affects the objectivity of the evaluation.

The central tendency effect refers to the tendency to avoid extreme scores by clustering around a mean. However, in this case, this bias does not apply to the first teacher, as evidenced by the highly dispersed scores, which refutes the hypothesis of a systematic central bias. Conversely, Teacher 2 illustrates this phenomenon, with a limited range of scores (minimum = 4.75, maximum = 16.5), reflecting a more cautious approach and a high concentration of evaluations around intermediate values.

Marker fatigue and variability in the criteria applied during grading are also potential sources of bias. Grading a large number of papers can lead to a loss of attention, excessive standardization, or, conversely, increasing severity. Furthermore, even when graders use a common rubric, the interpretation of the criteria remains subject to variation. Some prioritize grammatical quality, while others value the richness of ideas or the logical consistency of the argument. This variety of priorities reflects a degree of subjectivity which is part of human grading.

A bias towards leniency or strictness is clear in the results. Teacher 1's grades show that he is more generous, with a higher score, while Teacher 2 appears to be more rigorous in certain areas. This difference in grading approach is common, particularly in written assignments, where there is significant room for interpretation and where evaluation often varies according to the grader's personal style.

The halo effect was not observed here; it was neutralized in this study through the anonymization of the test papers.

It can be inferred that AI reduces certain human biases—notably the contamination effect, the contrast effect, and fatigue—by enabling rapid, standardized grading that significantly lightens teachers' workloads. However, it has limitations in evaluating written work, as it lacks the sensitivity and subjectivity necessary for a comprehensive pedagogical assessment. This aligns with the study by (Veldkamp and Sluijter, 2019). AI focuses on formal criteria (grammar, spelling, syntax) but struggles to assess writing style, argumentative coherence, and the student's communicative intent. It thus tends to apply fixed rules, leading to stricter grading that fails to value creativity, linguistic risk-taking, or efforts at rephrasing. This observation is particularly problematic in a context such as Morocco's, where French is a foreign language and where teachers often seek to encourage experimentation and interaction with the language.

Although AI offers benefits in terms of objectivity and consistency, it remains subject to its own biases, as confirmed by the study by Bulut et al. (2024), which identified several sources of bias: historical data bias, representation bias, measurement bias, aggregation bias, learning or algorithmic bias, and biases related to content generated by LLMs. In the context of this study, three forms of bias stand out in particular:

Learning or algorithmic bias: namely, a systematic performance discrepancy (OECD, 2024). In our study, an AI trained on academic texts tends to favor a rigid style and to place greater weight on formal criteria than on reasoning or argumentative logic. Consequently, an essay that is well-structured in terms of ideas but contains grammatical errors risks being

graded harshly. Similarly, original, creative, or personal writing may be undervalued, which puts certain students at a disadvantage, particularly in contexts where French is a foreign language, a finding consistent with (OECD, 2024, pp. 276–277).

Representation bias: Many AI models are trained primarily on English-language corpora, which can limit their ability to accurately assess certain structures specific to French, particularly in the variety used by non-native speakers. This cultural and linguistic influence limits the recognition of lexical, stylistic, or dialectal richness, introducing an additional form of inequality into the assessment.

Pedagogical biases: AI does not always take into account the pedagogical context, the teacher's intentions, the learner's actual proficiency level, or the specific objectives of the task. It also risks overlooking certain exceptions, missing important nuances, or ignoring the pragmatic intentions that guide the learning activity; thus, it may penalize a form even if it was expected in a free-style exercise.

AI is a promising complementary tool in educational assessment, capable of reducing certain human biases and enhancing consistency, but its own biases and limitations show that it cannot replace a teacher's interpretive sensitivity and pedagogical flexibility.

Conclusion

This study reveals a deep epistemological tension at the heart of the act of assessment. It proposed examining the reliability and objectivity of French language test grading by comparing two approaches: traditional human and AI-assisted grading. The main objective is to identify assessment biases specific to each of the two methods and to evaluate the extent to which AI could serve as an alternative to limit human subjectivity. The results revealed significant differences among human graders, particularly in terms of grading severity and score dispersion. The moderate level of correlation observed among evaluators confirmed inter-rater variability, exacerbated by the influence of biases such as central tendency, tiredness, or variability in the criteria applied. These findings show that each teacher, consciously or unconsciously, applies personal principles that lead to differences in marking. Conversely, AI-assisted grading stood out for its standardization of scores, lower score dispersion, and consistent application of criteria, demonstrating its potential to reduce certain human cognitive biases and increase the reliability of assessments. However, it has also revealed its own limitations, notably a difficulty in interpreting linguistic nuances, grasping the overall context, and assessing the creativity, richness of expression, or subtlety of style in written work. These subjective dimensions, which are essential to language learning, must not be entirely ignored. On the contrary, they must be managed and balanced to ensure an assessment that is fair, equitable, and educational.

Based on these findings, several avenues for improvement can be considered. Given the rapid evolution of AI, this would involve developing algorithms capable of analyzing the coherence of arguments and the structuring of ideas, better integrating linguistic and stylistic nuances, and recognizing the originality of student work. In addition, raising teachers' awareness of the existence of cognitive biases remains necessary to promote a hybrid approach, one that accepts that algorithmic rigor can free teachers from mechanical tasks, allowing them to refocus on their essential mission: clinical judgment and educational guidance. While far from replacing human judgment, the thoughtful integration of AI into educational assessment presents an opportunity to jointly develop a more equitable evaluation system capable of recognizing the uniqueness of each learner while ensuring equality. It can become a complementary assistant, allowing for a combination of algorithmic

rigor and pedagogical sensitivity, as long as research continues to scientifically validate its reliability and educational value.

References

- Bressoux, P., Dompnier, B., & Pansu, P. (2011). School Assessment: A Multifactorial Activity. In *Assessment: A Threat?*
- Bressoux, P., & Pansu, P. (2003). *When Teachers Assess Their Students*. Presses universitaires de France.
- Bulut, O., et al. (2024). The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges. [arXiv. 10.48550/arXiv.2406.18900](https://arxiv.org/abs/10.48550/arXiv.2406.18900)
- CNESCO. (2023). *School Assessment and Grading: Expert Reports*. Paris: National Council for the Evaluation of the School System. https://www.cnesco.fr/wp-content/uploads/2024/07/Cnesco-CC-Eval_Notes-des-experts.pdf
- De Ketele, J.-M., & Gerard, F.-M. (2005). Validation of Assessment Tests Using a Competency-Based Approach. *Measurement and Evaluation in Education*, 28(3), 1.
- De Ketele, J.-M. (2011). *Assessment and the Curriculum: Conceptual Foundations, Debates, and Issues*. *Educational Science Files*, 25, 89–106.
- Gobrecht et al. (2024). Beyond human subjectivity and error: A novel AI grading system. [arXiv](https://arxiv.org/abs/2406.18900).
- Gong, W. (2022). Reshaping the EFL Formative Assessment Pedagogy with Blockchain Technology. *International Journal of English Linguistics*, 13(1), 12. <https://doi.org/10.5539/ijel.v13n1p12>
- Guimard, P. (2010). *The Assessment of Academic Skills*. Presses universitaires de Rennes.
- Kolen, M. J., & Brennan, R. L. (2014). Linking. In *Test Equating, Scaling, and Linking: Methods and Practices* (pp. 487–536). New York, NY: Springer New York.
- McClelland, T., & Cuevas, J. (2020). A comparison of computer-based testing and paper-and-pencil testing in mathematics assessment. *The Online Journal of New Horizons in Education*, 10(2), 78–89.
- Mohammed Fatihi. (1995). *Measurement and Assessment Methods*. Didactica, Maroc
- OECD. (2023). *Teacher's Guide to Summative Classroom Assessment of Students in Morocco*. OECD Publishing. <https://doi.org/10.1787/2ac20f43-fr>
- OECD. (2024). Digital Assessment. In *OECD PERSPECTIVES ON DIGITAL EDUCATION: Toward an Effective Digital Ecosystem*; (OECD, pp. 126–149).
- Philippe Perrenoud. (1998). *Student Assessment: From the Creation of Excellence to the Regulation of Learning*. De Boeck Université.
- Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online Assessment in Mathematics and Writing: Reports from the NAEP Technology-Based Assessment Project, Research and Development Series*. NCES 2005-457. National Center for Education Statistics.
- Veldkamp, B. P., & Verschoor, A. J. (2019). Robust computerized adaptive testing. In *Theoretical and practical advances in computer-based educational measurement* (pp. 291–305). Cham: Springer International Publishing