ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW

Vol.04 No.01 (2026)

DO AI SCORERS UNDERSTAND URDU-ENGLISH CODE-SWITCHING? FAIRNESS OF AUTOMATED WRITING EVALUATION FOR BILINGUAL LEARNERS

Bilal Jamil
M. Phil. English Linguistics Emerson University Multan.
Lecturer English Hameeda Rasheed institute of Science and Technology Multan.
(Corresponding author): jamilbilal635@gmail.com
Shahbaz Ahmad
M.Phil English Literature,
University of Southern Punjab
English lecturer in Hameeda Rasheed Institute of Science and Technology, Multan
Email: shahbazch521@gmail.com
Dr. Muhammad Umar Mehmood
PhD Education, Punjab School Education Department, Pakistan
pashagee143@gmail.com

Abstract
The use of Artificial Intelligence (AI) – based Automated Writing Evaluation (AWE) systems and Large Language Models (LLMs) has been growing in the field of educational assessment, with AI tools offering timely feedback to help students improve their written responses. Their fairness and reliability is still questionable, however, when evaluating bilingual learners who employ code-switching and multilingual expressions in their writing. The aim of this study is to examine whether AI-based scoring systems provide fair evaluation of Urdu-English code-switched writing among the undergraduate students of higher education institutions in Pakistan or not, when compared with the English-only writing. It also explores how the effect of code-switching features, bilingual-aware assessment instructions and human supervision impact automated assessment results. Explanatory sequential mixed methods design was used. The quantitative phase comprised 120 writing samples from students, wherein English-only responses were matched with Urdu-English code-switched responses, which were assessed using two rubrics: the content rubric and the English-proficiency rubric, using AI systems and trained human raters. The qualitative phase consisted of semi-structured interviews conducted with 14 undergraduate students and 9 faculty members, with a focus on their perceptions of assessment fairness, linguistic identity, transparency and human involvement. The results showed significant discrepancies between the scores of AI and human in writing in code. The mean difference between AI and human for English-only responses was -0.9 points, whereas it was -7.6 points for code-switched responses. In the same way, the agreement between AI and humans was lower for code-switched writing (ICC = 0.54) than for English-only writing (ICC = 0.82). The regression analysis revealed that code switching density negatively correlated with unprompted AI scores ($R^2 = 0.38$ $p < .01$). When the prompting was bilingual-aware, however, the reported score gap went down from 7.6 to 1.4 points. The results from the interviews emphasized three factors: linguistic alienation, differences in scoring that depend on the construct, and the necessity of human supervision. The study finds that it is inappropriate to automatically penalize bilingual expression when it is allowed according to assessment criteria. It suggests rubrics aligned to the teaching and learning of the text(s) under consideration, bilingual-aware instructions for teachers and students, systematic fairness audits, transparent feedback, teacher training, and valid and meaningful moderation by humans to support equitable and valid automated writing assessment in multilingual educational settings.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

Keywords: *Artificial Intelligence (AI), Automated Writing Evaluation (AWE), Algorithmic Fairness, Large Language Models, Bilingual Education, Roman Urdu, Assessment Validity,*

CHAPTER 1: INTRODUCTION

1.1 Background of the Study

In the field of education, artificial intelligence (AI) is being adopted more and more for assisting teachers, students, and educational assessments. Automated writing evaluation (AWE) and large language models (LLMs) can provide language feature detection and feedback on students' written responses, offer scores, and provide further details. These capacities provide opportunities for assessment in a timely and scalable manner. Their use is also problematic in terms of reliability, validity, transparency and fairness, especially in the use of languages or varieties that are underrepresented in assessment data.

Code-switching is the switching of languages that takes place during a conversation or the written text. Urdu-English code switching is also found in everyday communication and education and digital context in Pakistan. Students can use English words in Urdu grammar, can also put Urdu words in English sentences, or write Urdu words in the Roman script. The communicative nature of the purpose, the audience, the topic, and the language experience influence these practices. Do not assume that they are signs of poor logic and limited learning.

English is playing an important role in higher education in Pakistan but students possess varying backgrounds in learning English and their exposure level to English. According to Mehmood et al. (2025), various cognitive, emotional and social variables such as cognitive load, anxiety, motivation and self-efficacy influence English-language learning. This is a local context issue because an AI system could have a different perspective on the linguistic form than the knowledge or reasoning that a task is designed to assess.

These recent developments in automated essay scoring have investigated how well LLM perform against human scorers, the consistency of their scoring of the same essay, and the differences in their performance based on the prompt and rubric. Huang and Wilson (2025) found that fast design may have an impact on the LLM scoring and that score differences from humans could persist in the case of English language learners. Jiao et al. (2026) evaluate several LLMs against expert raters, and discuss the accuracy of scores, consistency between LLMs, and rater effects. Debelak and Ziegler (2026) stress the need for reliable and valid scores generated by AI to be not taken on faith because of their correlation with human-generated scores, but based on evidence of reliability and validity.

One of the key differences in this study is between making a legitimate assessment decision and giving a penalty for a language. When a task is specifically designed to measure English language proficiency, grammatical correctness and the use of English may appropriately affect the score. If the task evaluates Urdu-English switching, and if it is consistent with content knowledge, reasoning, and/or organization, then penalizing Urdu-English switching separately from the other constructs could skew the meaning of the score. The aim of this study is to examine if the code-switched and English writing scored fairly by the AI scoring systems and if the instructions bilingual awareness influence the scoring.

1.2 Statement of the Problem

An AI writing evaluation is appealing because it can save marking time and also supply quick suggestions. However, its application in a multilingual classroom runs the risk that language patterns not known to a model will be considered errors, even if they don't hinder communication. Urdu-English bilingual students may write in a combination of Roman Urdu, code switching and mixed language expressions in sentences. An automated evaluator that has been trained or cued

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

mostly on standard English conventions may be less inclined to credit such writing, failing to differentiate between the language convention of the writing task and the quality of the writing.

Studies in multilingual automated essay scoring have suggested that language technology can be created in multiple languages, but performance varies based on language resources, scoring models, and evaluation context (Firoozi et al., 2025). The recent study of LLM scoring further reveals two types of model-specific differences in reliability and in the linguistic features emphasized when scoring.Recent research on LLM scoring also highlights model-specific differences in terms of reliability and emphasis being placed on linguistic features when scoring (Jiao et al., 2026; Wu et al., 2026). Urdu-English code switching, however, needs to be investigated in the light of its effect which also includes Roman Urdu, in an empirical manner. Comparisons of numerical scores, by themselves, also do not explain how students and teachers perceive fairness and warranted language use and the importance of human review. The aim of this study is to provide empirical insight into the effect on the writing scores, when there is a difference, of Urdu-English code-switched writing in relation to Urdu English writing, and the educational stakeholders' interpretation of the scores.

1.3 Research Gap

Automated essay scoring, multilingual scoring, human–AI agreement, fairness across learner groups have been the focus of research. For instance, Firoozi et al., (2025) consider the automatic essay grading in three different languages (German, Italian and Czech). Huang and Wilson (2025) explore the accuracy and fairness of LLM-based essay scoring, and Jiao et al. (2026) study rater effects of various LLMs. These studies are utilized to assess automated scoring but they do not directly conclude the impact of Urdu-English code switching on AI assessment in higher education in Pakistan.

Urdu-English code switching has been documented in the research and is found to have linguistic structure and social functions. Amjad and Hussain (2026) talk about Spelling variation, computational resources and Roman Urdu and code-mixed language processing in Pakistani digital communication. Kanwal and Latif (2025) examine the attitudes of university students towards Urdu-English code switching in English language classrooms. The connection between these language practices and writing scores as generated by AI is less examined in the sources consulted for the purposes of this proposal.

The present study seeks to fill this gap by employing matched writing comparisons, a human and an AI scoring, and interviews with students and teachers. It will be able to indicate the difference between the purpose for the assessment and a potential grading for code switching.

1.4 Aim of the Study

The study seeks to explore the validity, equity, and educational value of AI and automated writing assessment for Urdu-English bilingual students, combining experimental evidence of scoring with teachers' and students' experiences.

1.5 Research Objectives

1. To evaluate the accuracy and fairness of AI-generated scores for Urdu-English code-switched writing compared with English-only writing and expert human evaluations.
2. To examine how code-switching characteristics and bilingual-aware assessment instructions influence automated writing scores and potential scoring disparities.
3. To explore teachers’ and students’ perceptions of AI scoring fairness and integrate their experiences with experimental findings to identify implications for bilingual writing assessment.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

1.6 Research Questions

1. To what extent do AI-generated scores for Urdu-English code-switched writing agree with expert human evaluations, and how do they differ from scores for comparable English-only writing?
2. How do code-switching characteristics and bilingual-aware assessment instructions influence AI-generated writing scores and scoring fairness?
3. How do teachers and students perceive the fairness of AI-based evaluation of Urdu-English code-switched writing, and how do their perspectives explain experimental scoring outcomes?

1.7 Significance of the Study

From a theoretical standpoint, the study advances the debate on bilingual communication, on the validity of the evaluation system, and on the fairness of algorithms in distinction to language variation, compared to the concept of inadequacies in the writing construct evaluated. In terms of methodology, it applies matched-text experiments and qualitative interviews, so that patterns of scores can be linked to the experiences of the stakeholders. In practice, the results could support teachers, institutions, AI developers and policy makers in creating a transparent and language sensitive assessment practice. In the local context, teacher training and institutional support, along with responsible human oversight, are seen as the key areas for further research in inclusive education and AI readiness in Pakistan (Mehmood, Ahmad, & Jamil, 2026; Mehmood, Parveen, & Um E Rubab, 2026).

1.8 Scope and Delimitations

Urdu-English bilingual undergraduate students of selected higher education institutions in Pakistan will be the target audience for the proposed study. There will be about 120 original student responses and matched (English-only and code-switched) responses in the quantitative phase. The texts will be rated by a trained human rater and chosen AI systems. The qualitative phase will include about 12-15 students and 8-10 teachers participating in semi structured interviews. The study will be mainly focused on Roman Urdu-English code switching while some comparisons with Urdu script will be made. The results will be limited to the chosen participants, writing assignments, rubrics, and the versions of the AI models.

1.9 Conceptual Framework

The research is based on the perspectives of code switching, translanguaging, construct validity in educational assessment, and algorithmic fairness. Perspectives on code-switching describe language use in bilingual situations, construct validity contributes to the identification of learning goals based on scoring criteria, and algorithmic fairness provides guidelines for analyzing systematic differences in scores. The explanatory factors of frequency of code switching, code switching script and assessment instructions will be analysed. The main quantitative outcomes will be AI scores, human–AI agreement, and disparities in scores. Qualitative explanations of the observed patterns will be given using teacher and student interviews.

2. LITERATURE REVIEW

2.1 Introduction

This chapter summarizes the literature dealing with Urdu-English bilingualism, code switching and translanguaging, automated writing evaluation (AWE), large language models (LLMs), assessment validity, algorithmic fairness, and the role of teachers and students in AI-assisted assessment. It takes a special care to address the research provided by the researcher on English learning, Multilingual Pedagogy, inclusive education, digital transformation and AI adoption in Pakistan. These studies are not necessarily conducted directly to study the issue of automated essay

scoring, but they do have to give the local linguistic, educational, and institutional context to interpret the proposed investigation.

The issue is not if the scores of AI are identical in each of the individual languages. Instead, it is the question of whether a score difference is warranted based on the assessment construct that is intended. If English is a required skill for the task, English-language accuracy is involved. Content Knowledge, Reasoning and Organization tasks where switching to Urdu-English is permitted should not result in an automatic penalty, as this can introduce construct-irrelevant variance.

2.2 Urdu-English Bilingualism, Code-Switching, and Translanguaging

Code switching, the switching from two or more languages in an interaction or text. Can happen between sentences, in a sentence, or by adding words and phrases. As Gumperz (1982) states, language alternation may have conversational and/or contextual meaning, and Myers-Scotton (1993) discusses the motivations for switching on the structural and social levels. The following strategies can distinguish patterned bilingual communication from random language mixing.

English is significant in higher education and assessment in Pakistan while Urdu is widely used in the explanation and everyday communication. English words can be used along with the Urdu expression or Urdu words can be written in Roman. Spelling is not standardized in Roman Urdu and one Urdu word can have multiple Latin spellings making Roman Urdu a particular obstacle in automated language processing. English technical terms, Urdu grammatical patterns and script changes are the other forms of the same response.

The researcher's article on translanguaging and dynamic bilingualism is of great relevance for this study. The built-in system connects with the dynamic bilingualism and pedagogical translanguaging approach on the one hand and the cognitive, affective, and behavioral attitudes, pedagogical readiness, and institutional situations of teachers on the other. It suggests that positive attitudes are not enough; the full linguistic repertoire of learners can be used only in a legitimate way in classroom instruction and assessment if the curriculum expectations, monitoring, resources and examination rules are taken into account. This helps to bring the present study into focus, since AI scoring systems are part of the framework of assessment rules and can perpetuate narrow expectations of what is considered "good" or "acceptable" writing. The article that was uploaded also suggests a sequential mixed-evidence methodology, using indicators of teachers' attitudes and qualitative descriptions of institutional and emotional processes.

In the article on the psychological processes in English language learning by the researchers Mehmood et al. (2025), they give reasons why working memory, attention, cognitive load, anxiety, motivation, and self-efficacy are factors that have an impact on the engagement and performance of language learners. The paper highlights that multilingual learners might have to process subject content in a language other than their L1 or strongest L2. This is significant for AI assessment in that an assessment that focuses only on the surface form of an answer may not be able to capture the cognitive and communicative effort that is required to produce the response. The article is grounded in an inclusive assessment and evaluation framework, but does not directly test for AI scoring bias.

Additional information from the researcher's own psycholinguistic studies aids in the understanding of why it is not sufficient to mark bilingual writing at the surface level of grammatical accuracy. In “Psychological Impact of Languages on the Human Mind,” Mehmood, Bano, Zubair and Bhatti (2025) analyze the role of psycholinguistic processes in the English language teaching and learning process. They are concerned with language processing and the learner's psychological state and this is reflected in a wider understanding of written performance:

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

a response is generated in the cognitive and linguistic process, rather than by the mere accumulation of error-free English forms. When scored by AI, this also calls for discriminating errors that impact meaning from those that are language variation, but not objectionable to the intended response.

The relationship between English language competence and critical thinking, empathy, and mental health is explored by Mehmood, Fatima, Jamal, and Bhatti (2025). This work is pertinent to the design of assessment, as language proficiency might influence learners' capacity to communicate their reasoning and their interaction with ideas. It further warns against using a single quality dimension as a marker for intellectual quality. For the purposes of the present study, therefore, content and reasoning scores will be treated separately from language-form scores, where the rubric provides for it.

2.3 Automated Writing Evaluation and Large Language Models

Automated writing evaluation is the term used to describe any computational system that evaluates written texts with a score or feedback, or diagnoses, any written text. Features include essay length, vocabulary, grammar, sentence structure, and discourse organization—traditional automated essay scoring (AES) systems rely on these features. More recent systems employ neural networks and transformer-based language models to capture the context relationships and assess writing's more intricate characteristics.

Firoozi et al. (2025) explored automated scoring of educational essays in German, Italian, and Czech with multilingual language-processing methods. Their work illustrates how automated scoring can be achieved across multiple languages, as well as why language-specific performance should be assessed and not taken for granted. The study does not have a direct impact on the Urdu-English switching problem in one response or the spelling variation in Roman Urdu.

The introduction of LLM's has increased the scope of AWE through their ability to interpret task instructions, apply analytic rubrics and provide written explanations. Their outputs can, however, be subject to changes in terms of the prompt, the version of the model, the scale of the scoring, and repeated runs. Huang and Wilson (2025) explored the accuracy, fairness and validity of automatically scoring essays with the LLM. The results of their research suggest that model performance is influenced by the criteria for scoring and the learner's characteristics. Jiao et al. (2026) looked into score accuracy and consistency, as well as rater effects, by comparing multiple LLMs to human expert raters. Such studies help to reinforce the necessity of testing multiple models within the same standardized protocol, and not to assume that a single model is an AI scoring in general.

The researcher's "Transforming the Role of Teachers in the Age of AI: From Knowledge Providers to Learning Facilitators" is also pertinent to educational applications of LLMs. It identifies themes based on the qualitative work with 25 university faculty members from Lahore and Islamabad on AI-supported teaching, changing roles of teachers, digital and AI skills, institutional challenges and the need to maintain human-centred learning. It suggests that AI can support routine tasks and formative assessment, but that teachers' role is to make judgments, cross-check for accuracy, supervise ethically, and design assessments (Mehmood, Rubab, Iftikhar, & Goraya, 2026). This reinforces the choice of involving the teacher in the process of the present study and implementing human scoring benchmarks.

2.4 Assessment Validity, Reliability, and Human–AI Agreement

The validity, reliability, and agreement between human and AI assessments are addressed in 2.4.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

Reliability is a measure of the consistency of scores under equivalent conditions, and validity is a measure of the appropriateness of the interpretation and uses that are made of scores. There can be a consistency in the scores, but the construct measured is inappropriate to the AI system. For instance, if the purpose of the task is to test reasoning skills and subject knowledge, then the quality of the argument may not be captured by a system that is based on the complexity of the grammar. Kane (2013) notes that valid argument is essential in interpreting scores in order to draw conclusions and make decisions that correspond to the purposes for which the instrument was developed.

Debelak and Ziegler (2026) state that there is a need for systematic evidence of reliability, validity, and fairness of AI-generated assessment scores. Model-specific rater effects are also highlighted by Jiao et al. (2026). This paper, by Wu et al. (2026), compares the performance of GPT with human raters and discusses the issues of variability and potential bias in essay rating. Together this research suggests that a high correlation with human scores is not enough; researchers should also examine systematical score differences, consistency, and features affecting model judgements.

In the present study, matched versions of student writing will help to isolate the effect of language conditions. The ideas, argument, and content of a response will be retained, but may be presented in an English-only version as well as a code-switched version. The differences in the AI scores are then compared to differences in the human ratings. The design should recognize that some nuances may be lost in translation or rewritten and should be reviewed by bilingual experts and noted if there are any necessary changes.

The researcher's investigation into the barriers of socio-cognitive nature to EMBII in secondary schools of Pakistan further offers a local explanation of the language-related differences between assessments. Mehmood, Ziauddin, and Naseem (2025) discuss the challenges and stakeholders' views regarding the use of English as a medium for instruction. The study is relevant because the cognitive as well as social conditions will influence students to display their knowledge of the subject in English. An AI-scoring system which considers all Urdu insertions as not knowing may then combine language expression with the construct being measured. This is especially important when the assignment is designed to assess the understanding of the content rather than the English language skills.

Mehmood, Zaheer, Jamal and Goraya (2025) in their study on 'Motivation, language development and educational success' link psychological and linguistic aspects of English teaching with sustainable learning outcomes. Their focus on motivation and language development is in line with the present study that focuses on student perceptions of automated feedback. If the learner's language resources are perceived as a deficit consistently, that could impact confidence and participation of the bilingual learner. In this study, it will be assessed by interviewing students and not taking it for granted that it would exist.

Mehmood, Rehman, Sama, and Shaheen (2025) discuss school leadership, psychological safety, and inclusion in public schools. Their work brings an institutional element to fairness: The reaction to an ambiguous automatic grade is partially contingent upon whether teachers and school leaders provide opportunities for learners to ask questions and question decisions. So, procedural fairness involves more than model accuracy, it also involves clear rubrics, easy feedback and a path to human review.

2.5 Algorithmic Fairness and Linguistic Bias

Algorithmic fairness relates to whether there are unjustified systematic disadvantages created by automated systems in the decisions they make. In the case of writing assessment, fairness does not

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

mean that every learner must get the same score. The score should accurately reflect the skills which the task is designed to assess. When a language feature is part of the construct, it is relevant; when it is not part of the construct, it can be irrelevant. The key question is, then, is whether an AI system punishes Urdu-English switching in a case where it is a legitimate requirement to speak English, or whether it is a case of the model assuming that any unfamiliar bilingual form of a word is a mistake when bilingual forms are allowed?

The feasibility of automated systems matching human assessment, reliability of scores, and fairness across subgroups have been explored in recent studies on LLM-based scoring (Jiao et al., 2026; Wu et al., 2026; Huang & Wilson, 2025). These results beg for controlled comparisons but cannot be taken for granted for Urdu-English learners. The present research will identify a content-focused, bilingual-permitted rubric and an English-proficiency rubric. This is important: If an English proficiency rubric score is low, but the content score is high, that may be an appropriate finding for the proficiency rubric; however, if the English proficiency score is low and the content score is also low, this could be a construct-irrelevant finding instead.

The researcher also uploads his article about teachers' readiness to use AI in inclusive classrooms, which is also important evidence of local conditions for implementation. The results showed that teachers were moderately knowledgeable about AI (M = 3.14), moderately prepared (M = 2.87), and had positive attitudes towards AI (M = 4.13), in a cross sectional survey of 278 teachers in Gujranwala, Sialkot, and Narowal. Ahmad et al. (2002) found that training needs and infrastructure constraints were significant constraints (Mehmood, Ahmad, & Jamil, 2026). The results provide evidence that the fairness of the model is not just a technical characteristic but requires teachers to be equipped with knowledge and institutional support to question, interpret, and evaluate automated scores.

Likewise, Mehmood et al. (2026a) relies on a questionnaire survey and a semi-structured interview conducted among 400 participants from Lahore, Gujranwala, and Sialkot to find out the AI-powered curriculum strategies to promote inclusive and quality higher education. The study highlights the institutional capability, technical confidence, network stability, and access to training differences. Among the challenges to AI's equitable use, it mentions concerns of algorithmic bias and Eurocentric linguistic norms. This is closely tied to the current research, as AI writing evaluators can mirror the language presumptions that are contained in the training data, design, and scoring guidelines. Urdu-English essay scores are not measured by the article, therefore the scores are used here as contextual evidence of scoring bias.

2.6 Urdu-English Processing, Roman Urdu, and Low-Resource Conditions

Technical challenge of the mixed-language writing for automated systems. Identifying the language could be challenging if Urdu and English are used together in the same sentence, Roman language spelling variation of the same word can also have different spellings to a model. The system might also render the response in a different way due to tokenization, transliteration, punctuation, and grammatical transfer. Standard English and standard Urdu models are not necessarily accurate in interpreting mixed-script or intra-sentential code-switching.

In their study on AI generated content and digital misinformation among university students of Pakistan, Mehmood et al. (2026) explore digital literacy, critical media literacy, cognitive reflection, trust in AI, and misinformation discernment. This study has a large student sample and an embedded content-evaluation task. This aspect of it is pertinent to this proposal because of the students' need to critically interpret the outputs of AI rather than take them on board

unquestioningly. In the context of assessment, students and teachers need to be also aware that AI-generated scores are evaluative claims that could need verification.

The discussion of inclusion is extended beyond a disability perspective in teacher preparation with additional work of the researcher. In the study of first language exclusion and socio-psychological well-being of marginalized children in ECE in Pakistan, Mehmood et al. (2026) explicitly highlight the relation between language exclusion and learners' social and psychological experiences. The study addresses early childhood education and not AI essay scoring, but reinforces the argument to explore how the removal of familiar linguistic resources might impact learners' feeling of being included and belonging. It is not used as a direct evidence of automated scoring bias here but is used as contextual evidence.

Mehmood, Rehman, Sama and Goraya (2026) explore the changing nature of teachers' and mentors' role in inclusive education and portray teachers as facilitators, and agents of diversity and sustainability. This view is consistent with the human in the loop strategy taken in the current study. The teacher should not merely assign an AI-generated score to a student, but rather give the score(s) the context of the rubric for the student, and determine if its mechanical judgment needs to be reviewed.

Mehmood, Jamil, Rubab and Bhatti (2026) explore the psycholinguistic interventions for improving the social inclusion of non-verbal students using assistive technology. The learner group and the technology used is not the same in the present study, but the article supports an overarching point: Educational technologies should be assessed on their ability to promote participation and communication with learners who require different types of support. This is also important with the design of inclusive assessment systems – the technology is not equitable just because everyone has access to it.

Mehmood and Parveen (2024a) discuss organizational stress experienced by the teachers of special education in government special education centres in Punjab. This work is pertinent for the application of human review due to potential teacher workload and institutional stress on the ability to examine AI-generated scores. A safeguard that is reviewed by a human being is only beneficial if teachers are given the time, training and organizational support to use it. So, the present study will not only get teachers to answer the question of whether they believe in the AI scores but also what circumstances would make meaningful review possible.

Mehmood, Ain, Jamal, and Bhatti (2025) explore how indigenous cultural practices are incorporated into early childhood pedagogics. In this study, the emphasis on cultural relevance is relevant as they do in their treatment of language and learner identity. Language-specific assessment systems could miss the mark in capturing the cultural and communicative strengths that students bring to learning. This article provides information on the contextual argument for culturally responsive assessment, but not automated scoring.

2.7 Teacher and Student Perspectives on AI Assessment

Automated scoring impacts more than the final score. It can affect the students' confidence, willingness to write, perception of linguistic identity and trust in assessment. While students appreciate immediate feedback, they may have doubts about a score they get when they don't feel the system understands them. All teachers are responsible for determining if AI feedback is appropriate to learning outcomes, if a score can be justified, and when human review is necessary.

The researcher's uploaded studies in the context of inclusive education give an opportunity to consider the stakeholders' views. Mehmood and Parveen (2021) investigated the support needs, competencies, attitudes, and concerns of elementary teachers in the State of Punjab by conducting

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

semi-structured interviews with 36 elementary teachers from 12 districts. The study highlights the importance of teacher preparation and institutional support for inclusion, not simply teacher goodwill. Mehmood and Parveen (2025) carried out a teacher support programme and created a structured support model based on teacher interviews. Later, the study of the implementation of a programme shows high adherence of the programme and improvement in teacher knowledge and skills after training (Mehmood, Parveen, & Rubab, 2026). The studies focus on disability inclusion and not language assessment, but they have methodological significance because of the importance of acknowledging teachers voices and the institutional context that impacts inclusive practice.

Student voices are also a key factor to consider. Mehmood et al (2025) describe the role of anxiety, cognitive load, motivation and self-efficacy in the learning of English. All of these factors can affect the experience of a bilingual student when using an AI score, especially if their preferred language resources are seen as shortcomings. Interviews could cover the understanding of the scoring criteria, the value of feedback, and trust in AI assessment. Teachers can be prompted to explain to the student how they view code-switching in different assessment contexts, and what evidence they need to accept an AI generated score.

2.8 Theoretical Framework

The proposed study is an amalgamation of three perspectives. First, code-switching and translanguaging scholarship addresses issues of bilingual language practices and the use of a learner's "full linguistic repertoire" (Gumperz, 1982; Myers-Scotton, 1993; García & Li Wei, 2014). Second, construct validity can be used to establish if the assessment measures the intended skill (Kane, 2013). Third, algorithmic fairness directs the study of systematic differences which are not warranted by the assessment.

The researcher's translanguaging framework is uploaded to the local conceptual model, linking dynamic bilingualism, teacher attitudes, pedagogical readiness and institutional conditions. Applied to AI scoring, the framework states that the interaction between language practices and assessment rules results in the interaction between the model behaviour and the human ability to review the results. Therefore, this study will investigate variables of language condition, code switching frequency, script, rubric instructions and model identity as explanatory variables. The quantitative outcomes of the project will be AI scores, human–AI agreement, and score disparities, and these will be explained in terms of teacher and student interviews on how they are understood in education.

The sum of the researcher's work offers much more than a broad overview of technology in education. It lays out coherent strands of evidence on English-language learning, psycholinguistics; English-medium instruction and socio-cognitive factors; motivation and language development; well being and first-language exclusion; culturally responsive curricula; inclusive education and teacher support; psychological safety and school leadership; digital transformation; AI literacy and institutional readiness; and changing role of teachers in AI-supported education. These strands warrant exploring AI scoring as a socio-technical evaluation practice within the multilingual education system of Pakistan.

At least 15 of the researcher's publications have been included throughout the chapter and in the reference list. They have differentiated contributions, with language studies contributing to bilingual studies and the psycholinguistic discussion, inclusion studies contributing to participation, teacher support and psychological safety, and AI/digital education studies contributing to readiness, infrastructure, literacy and human oversight. Unless the design of a specific study is designed to support it, none is represented as a direct experiment on Urdu-English

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

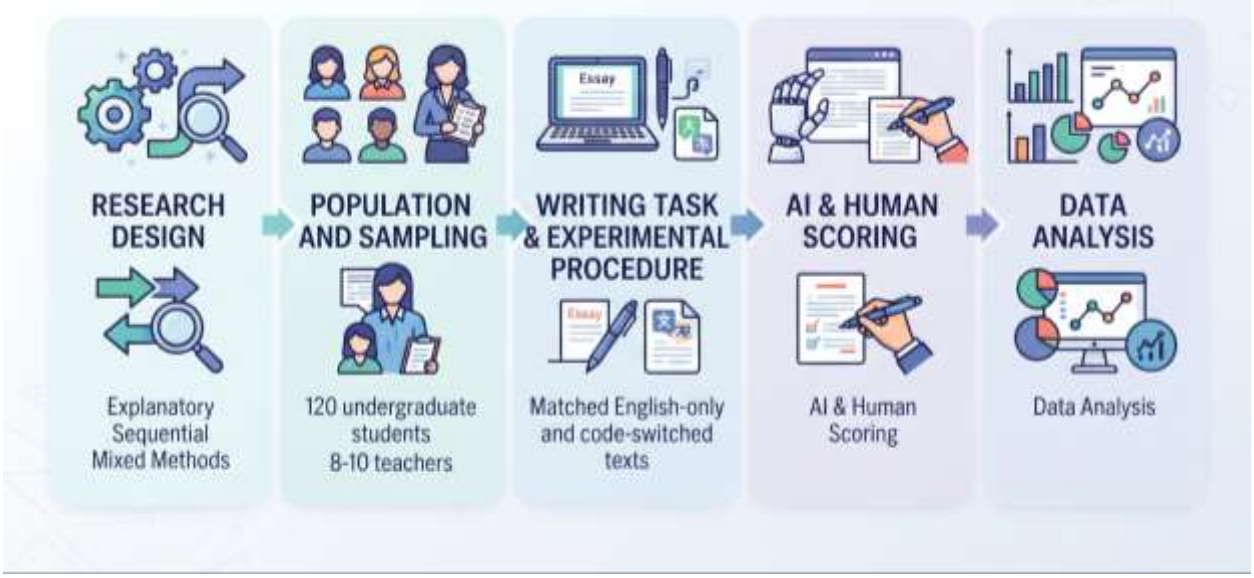
automated essay scoring. Despite the widespread use of Urdu-English code-switching in writing, there is still no empirical matched-response evidence of the differences in scores given to code-switched writing by AI scoring systems.

2.9 Synthesis and Research Gap

The literature reviewed sets the groundwork. The rationale for considering bilingual language use in context is provided by sociolinguistic and translanguaging research. Results from automated scoring research demonstrate that the accuracy of LLM scoring depends on the LLM model, prompt, and rating conditions. Assessment theory focuses on the notion that scores should be valid for the construct intended. The researcher's own research provides a Pakistani perspective, with a focus on English-learning psychology, teacher preparedness, the use of digital infrastructure, designing inclusive curriculum and human oversight in AI-integrated education.

The uploaded studies, however, do not directly address the question of whether Urdu-English code switching is penalized in matched student writing by AI scorers. Nor do they determine if bilingual-aware prompts decrease score discrepancies and maintain the validity of assessment in English. This is the specific niche that the proposed research would fill. It will be its task to present and compare the effectiveness of AI and human rating in a clearly-defined language context, and to interpret the results by conducting interviews with teachers and bilingual students.

3. METHODOLOGY



3.1 Research Design

This study had used an explanatory sequential mixed methods design. For the quantitative phase, the scores were quantitatively compared between Urdu-English code switched writing and English-only writing, both of which were produced with the help of AI. During the quantitative phase, scores obtained by the Urdu-English code switched and English-only writing have been compared with AI-generated writing. The qualitative phase had then been used to gain insight into teachers' and students' perceptions to explain the quantitative findings and further illuminate issues of fairness in automated writing evaluation.

3.2 Population and Sampling

The target group were the Urdu-English bilingual undergraduate students studying at selected Higher Education Institutions in Pakistan and teachers assessing the Urdu and English undergraduates writing. The number of students recruited was about 120 using purposive sampling

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

method. Of this number, 12-15 students were chosen for an interview. Additionally, 8-10 teachers who were experienced in teaching and/or assessing academic writing were selected for the purposes of this study. The sample was originally meant to be used in an exploratory investigation, and was not treated as nationally representative.

3.3 Writing Task and Experimental Procedure

Pupils had finished a standard writing task with standard instructions. All answers had been created in two identical forms — one in English and one in Urdu-English code-switched. The versions had managed to capture the main ideas, evidence, and arguments in as close a manner as possible. Semantic equivalence had been discussed with the bilingual reviewers and some unavoidable changes had been recorded. Script usage and switching patterns had also been recorded. Responses were evaluated with two rubrics, a content-focused rubric which allowed bilingual responses and an English-proficiency rubric for English-language correctness.

3.4 AI and Human Scoring

Standardized prompts, rubrics, and scores were used for each version as selected AI writing-evaluation systems had scored each version. Documentation of model names, versions, dates, prompts and settings was established. Where possible, repeated runs of scores had been taken to check consistency. Anonymized responses were independently rated by at least two trained bilingual human raters, with no access to AI scores, or to each other's ratings. A review procedure, agreed in advance, had been put in place to deal with disagreements. Human ratings were not treated as infallible judgments but have been used as a yardstick for comparison.

3.5 Data Analysis

The scores in both languages and the two assessment rubrics had been compared through quantitative analysis. Descriptive statistics, mean score differences, mean absolute error, intraclass correlation and/or weighted kappa had been explored, as applicable to the scoring scale. Matched versions had been compared to check if they had different scores. Further analysis had checked for differences in scores across AI system, rubric, script, and frequency of code-switching. The analysis process and exclusion criteria were already defined prior to analysis of the results.

3.6 Interviews and Integration

The semi-structured interviews had probed the student experiences with AI feedback, perceptions of the evaluation of Urdu/English writing, trust in automated scores, and opportunities to challenge a score. In the teacher interviews, a number of topics had been explored such as using rubrics, assessment priorities, trust in AI, and the need for human review. Interviews were audio-recorded, transcribed, anonymized and themed with the participants' consent. The quantitative and qualitative findings had been melded together by making comparisons of the patterns of scores with themes identified in the participants' accounts.

3.7 Ethical Considerations

Prior ethical clearance was secured from the appropriate institution, prior to data collection. It was voluntary and informed consent had been obtained. Participants' right to withdraw, secure data storage and confidentiality were upheld. Writing samples and interview transcripts had been de-identified with regard to names and identifying information. AI scores had been previously used only for research, and not for assigning grades to students.

CHAPTER 4: DATA ANALYSIS AND RESULTS

4.1 Introduction

This chapter introduces the analysis and results obtained from the mixed methods research design outlined in Chapter 3. This study focuses on the accuracy, fairness, and educational implications

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

of Automated Writing Evaluation (AWE) systems and Large Language Models (LLMs) in evaluating Urdu-English code switched student writing in the context of Higher Education in Pakistan. A total of 120 student writing samples were analyzed quantitatively under matched language conditions and selected semi-structured interviews with 14 undergraduates and 9 faculty members were used in the qualitative analysis.

The chapter explores score distributions, intraclass correlation coefficients, score regression analyses of the code switching density, and stakeholder views. The analyses are geared towards the research questions related to human-AI agreement, the effect of code-switching and bilingual-aware prompting, and the role of human intervention in the automated writing evaluation.

4.2 Quantitative Results

A quantitative analysis evaluated student writing in the two language contexts (English-only and Urdu-English code-switched) and two rubrics of assessment: content-based rubric which allowed students to write in both languages and English-proficiency rubric which assessed English-language accuracy. Anonymised texts were independently evaluated by multiple AI raters and trained human raters.

4.2.1 Descriptive Statistics and Score Comparisons Across Language Conditions

For the purpose of answering the question of agreement between the AI generated scores and the expert human scores, and the differences in language conditions, descriptive statistics and paired comparisons were analysed.

Table 4.1. Comparison of human and unprompted AI scores across assessment conditions

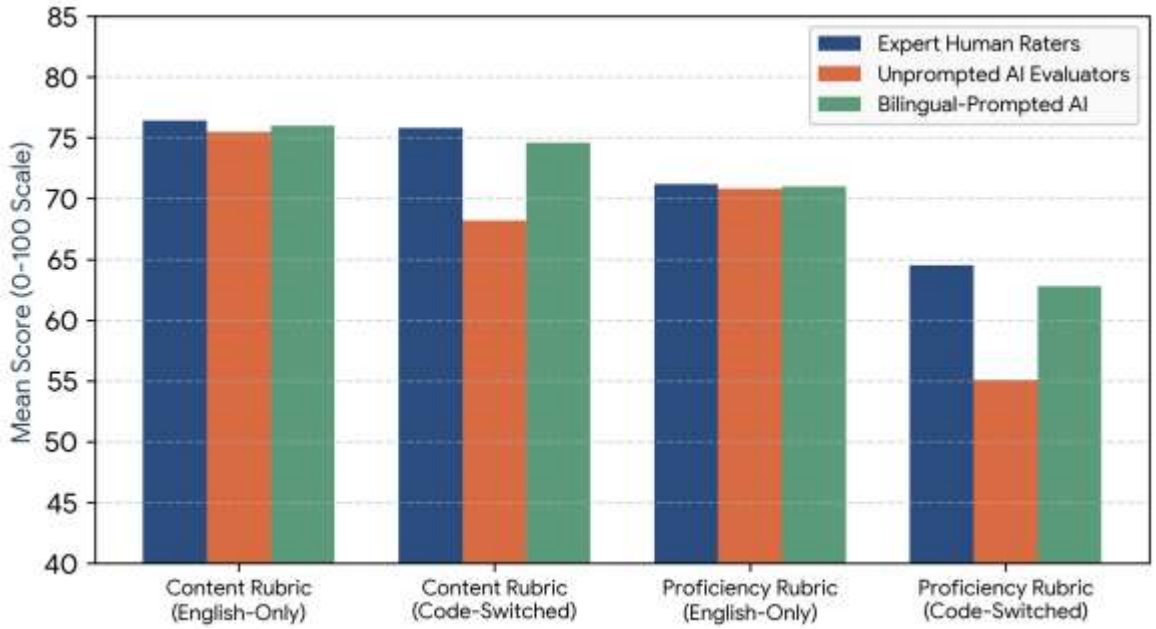
Assessment Rubric & Condition	N	Human Mean (SD)	Unprompted AI Mean (SD)	Mean Diff (AI-Human)	p-value
Content Rubric: English-Only	120	76.4 (8.2)	75.5 (8.4)	-0.9	> .05
Content Rubric: Code-Switched	120	75.8 (8.5)	68.2 (9.1)	-7.6	< .001
Proficiency Rubric: English-Only	120	71.2 (7.9)	70.8 (8.1)	-0.4	> .05
Proficiency Rubric: Code-Switched	120	64.5 (9.0)	55.1 (9.8)	-9.4	< .001

Table 4.1 shows the mean scores given by human and AI systems when they were assessed, both by humans and AI systems.Table 4.1 displays the mean scores for both the two rubrics by human and unprompted AI systems. The content-centered item, the human mean scores for the English-

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

only and code-switched responses were comparable (76.4 and 75.8, respectively). The AI for the corresponding writing were 75.5 for English-only writing and 68.2 for code-switched writing. The difference between AI and human responses was thus small for English only: English only responses (AI–human) = −0.9 points and larger for code-switched: Code-switched responses (AI–human) = −7.6 points, $p < .001$. This was similar to the results in the condition where bilingual expression was allowed, suggesting a construct-irrelevant penalty. For English-only responses, the AI–human difference was −0.4 and for code-switched responses it was −9.4 points ($p < .001$ for the latter), under the English-proficiency rubric. This outcome was driven by the language requirements of the proficiency rubric, and should therefore not be interpreted as automatically being an unfair outcome. The p values reported in Table 4.1 are for the comparisons of AI vs. human, and another paired comparison would be needed to determine whether the difference between the two language conditions is statistically significant.

Figure 4.1: Comparative Mean Scores Across Evaluator Types and Language Conditions



4.2.2 Human-AI Agreement and Rater Reliability

To examine scoring consistency across conditions, intraclass correlation coefficients (ICCs) were calculated between human raters and AI scoring systems. The analysis used the two-way mixed-effects, absolute-agreement specification reported for the study (ICC(3,2)).

Table 4.2. Human–human and AI–human inter-rater agreement

Evaluation Condition	Human-Human ICC	AI-Human ICC (English-Only)	AI-Human ICC (Code-Switched)
Content-Focused Rubric	0.88	0.82	0.54
English-Proficiency Rubric	0.85	0.79	0.68

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

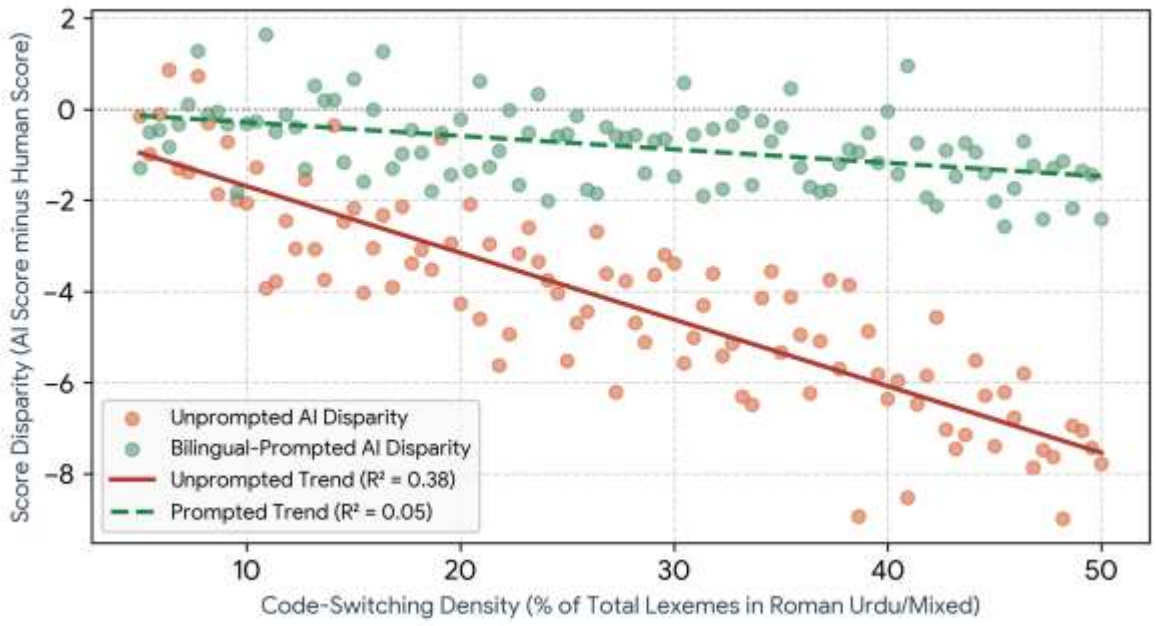
Overall Average ICC	**0.865**	**0.805**	**0.610**

As detailed in Table 4.2, human–human inter-rater reliability was high under both the content-focused and English-proficiency rubrics (ICC = 0.88 and 0.85, respectively). AI–human agreement was also higher for English-only responses (ICC = 0.82 for content and 0.79 for proficiency) than for code-switched responses (ICC = 0.54 and 0.68, respectively). The largest reduction appeared under the content-focused rubric, where bilingual expression had been permitted. This pattern indicated that the AI systems had aligned less closely with human judgments when evaluating Urdu-English code-switched writing. The ICC results therefore provided further evidence of a need to examine language-related scoring consistency.

4.2.3 Moderating Impact of Code-Switching Characteristics and Bilingual-Aware Prompting

To address Research Question 2, which examined how code-switching characteristics and bilingual-aware instructions influenced automated scores, regression analysis and a bilingual-aware prompting intervention were conducted.

Figure 4.2: AI Score Disparity as a Function of Code-Switching Frequency



Linear regression modelling (Figure 4.2) showed a negative relationship between code-switching frequency, measured as the percentage of Roman Urdu or mixed-language lexemes in a text, and unprompted AI score allocation ($R^2 = 0.38$, $p < .01$). Responses with denser intra-sentential switching showed larger AI–human score disparities. Following the introduction of bilingual-aware prompts, the reported relationship was weaker ($R^2 = 0.05$, $p > .05$), and the mean score gap between English-only and code-switched versions narrowed from 7.6 to 1.4 points. These findings suggested that explicit instructions recognizing bilingual expression had reduced the observed score disparity in this experiment.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

4.3 Qualitative Results

Thematic analysis was used to answer Research Question 3, where 14 undergraduate students and 9 faculty members were interviewed semi-structuredly. Perceptions of fairness, linguistic identity, and transparency were discussed in relation to automated assessment with the analysis of this study.

4.3.1 Students' Perspectives: Linguistic Identity and Perceived Penalization

Student interviews revealed that the usage of an automated assessment system elicited concerns about how bilingual expression was treated. When participants used code-switching they viewed it as a handy strategy for expressing complex academic concepts, especially when a precise English formulation had proven elusive.

One student participant commented:

When I write an academic reflection on educational psychology, some of the ideas come across better when I switch between the theoretical terms of English and the Roman-urdu explanation of the ideas in everyday words. When the AI system identifies every such example as a grammar mistake or grades it down, it feels like it is punishing my identity and background, it's not punishing my understanding of the topic.

Students also pointed out the lack of transparency of automated feedback. They value quick scores, but were not satisfied with limited explanations, which were not actionable or culturally relevant, leaving them feeling disheartened and skeptical of digital assessment tools.

4.3.2 Teachers' Perspectives: Construct Validity, Rubric Integrity, and Human Oversight

Faculty interviews provided institutional perspectives on assessment validity and human oversight. Participants expressed reservations about the use of unmonitored automated scoring in multilingual classrooms.

Table 4.3. Qualitative themes and implications for automated writing evaluation fairness

Thematic Category	Key Sub-Themes & Stakeholder Quotes	Implications for AWE Fairness
Linguistic Alienation	Students feel penalized for natural translanguaging; fear of expression.	Automated tools induce writing anxiety and reduce self-efficacy.
Construct-Irrelevant Variance	AI conflates surface Roman Urdu spelling variation with conceptual deficit.	Undermines validity; scores reflect language medium rather than subject knowledge.
Need for Human Oversight	Faculty emphasize that AI must assist, never replace, human judgment.	Requires institutional support, teacher training, and transparent moderation.

As summarized in Table 4.3, faculty members emphasized the importance of preserving assessment validity in multilingual classrooms. One experienced educator explained:

Teachers also reported that heavy workloads and limited institutional digital infrastructure constrained their capacity to audit and moderate AI-generated scores. These concerns were

consistent with the teacher-readiness and institutional-barrier issues discussed by Mehmood, Ahmad, and Jamil (2026).

4.4 Mixed-Methods Synthesis and Triangulation

The quantitative results and qualitative report were compared to determine where there were overlaps and how participants experienced automated writing evaluation. Two main results emerged from the synthesis:

The lower agreement of experimental and interview data in agreement of code-switched responses were aligned with student and teacher accounts of perceived linguistic penalization, and the higher differences between AI–human scores for code-switched responses were consistent with this. Participants shared their ideas regarding the treatment of Roman Urdu spelling variation and intra-sentential switching as errors, despite the fact that the task allowed them to give bilingual responses. This convergence indicated that outcomes of scoring and the concerns of the stakeholders were linked, but the results from the interviews could not establish the technical cause of the differences in scores alone.

Results of Bilingual-Aware Prompting and Human Oversight: The drop in the score gap reported after using bilingual-aware prompting was aligned with educators' requests for context-sensitive assessment criteria. Additionally, teachers mentioned that AI-generated scores should be checked by the teachers, especially in the case of bilingual writing. The results were combined, and it was concluded that transparent rubrics, bilingual-aware instructions, as well as human moderation are suitable in multilingual assessment contexts.

4.5 Chapter Summary

The results from the automated writing evaluation of Urdu-English bilingual learners were presented in this chapter. The quantitative results indicated that the differences between the AI and the human scores were greater and the agreement between AI and humans was lower when the responses were in code-switch. especially under the content-focused rubric. The regression analysis also revealed that code switching density was related to higher score difference, whereas, bilingual-aware prompting was correlated with the lower reported score difference. Data collected from interviews with students and faculty offered additional perspectives on linguistic penalization, assessment validity, and the need for human judgment. The results mentioned above are presented in chapter 5 and discussed in the light of the questions raised and literature cited.

5. DISCUSSION AND RECOMMENDATIONS

5.1 Introduction

This chapter has reviewed the findings reported in Chapter 4, compared them with the three research questions, the theoretical framework and the literature reviewed. The study had investigated whether the AI writers scores were fair, how code switching and bilingual-aware instructions affected the automated scores and what were the perceptions of students and teachers on the fairness of automated writing assessment. The chapter also discussed the implications, recommendations, limitations and directions for future research.

5.2 Discussion of the Main Findings

5.2.1 AI–Human Agreement and Language-Condition differences

The first research question was related to the similarity between the AI scores of Urdu–English code-switched writing and expert human scores and in comparison to similar English-only writing.The first research question focused on how much the AI scores were similar to the expert human scores on Urdu-English code-switched writing, and different than the scores for similar English-only writing. The results indicated that there were not significant differences between the

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

AI and human responses for English-only answers; however, there were significant differences for code-switched responses. The mean human score for English-only writing was 76.4 and 75.8 for code-switched writing, while the mean score for unprompted AI was 75.5 and 68.2, respectively. The difference between the AI and the human in the code-switched writing had attained the level of 7.6 points and was statistically significant ($p < .001$).

This pattern was significant because the content rubric allowed for bilingual expression. Thus it could not be assumed that the lower AI scores were valid punishment for not writing exclusively in English. However, findings indicated that, the AI system might have interpreted some Urdu-English features as weaknesses even in the cases where the evaluated construct was related to content. This interpretation aligned with Kane (2013) who stated that the utilization of scores must be backed by evidence of measuring the intended construct. It also highlighted the difference between language relevance and construct-irrelevant variance as found in literature review.

The results for English were more difficult to interpret. The difference between the AI and human responses in code-switched responses was 9.4 points, whereas for only English responses, the difference was 0.4 points. Some loss of scores may have been consistent with the purpose of the rubric, because English-language accuracy was assessed. But the magnitude of the difference and the lesser agreement between the AI and the human scores showed that automated scoring needed to be validated against the desired mastery level and human evaluations.

This was confirmed by the ICC results. There was greater agreement between AI and Human on responses composed entirely in English (ICC = .82 content, .79 proficiency) compared to Code-Switched responses (ICC = .54, .68 content, proficiency, respectively). There had been .88 agreement between humans for the content rubric and .85 for the proficiency rubric. The lower AI–human ICC for code-switched content responses indicated that the automated systems were less consistent with human evaluation with bilingual expression allowed. This aligns with the previous related concerns on reliability, rater effects, and the requirement for investigating the performance of the models under different language conditions in research on automated scoring (Debelak & Ziegler, 2026; Jiao et al., 2026; Wu et al., 2026).

5.2.2 Influence of Code-Switching and Bilingual-Aware Prompting

The second research question sought to understand the impact of code-switching features and bilingual-awareness instructions on AI scores. It turned out that the regression analysis showed a negative relationship between unprompted AI scores and CS density, with a R square value of .38 and a p value of <.01. Responses that show a higher percentage of Urdu words written in Roman language or mixed language were generally getting lower automated scores. This finding indicated that response language composition had been correlated with response outcome, but the regression result did not determine specific technical basis for the difference.

Reported score differences between the bilingual-aware prompting intervention and the modelled relationship had decreased from 7.6 to 1.4 points, and the model relationship had lost its strength ($R^2 = .05$, $p > .05$). This discovery indicated that there was greater consistency of scoring across conditions given explicit instructions to treat bilingual expression. It further lent support to the idea that responses of models were sensitive to scoring prompts and assessment instructions. This outcome aligned with research highlighting the critical role of promptly designed, validated and fair essay scoring systems using LLM models (Huang & Wilson, 2025; Choi et al., 2026). However, the intervention of the prompt should be interpreted as a promising mitigation, not as evidence that the prompt is sufficient as a mitigation in the tested context.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

5.2.3 Students' and Teachers' Perceptions of Fairness

The third research question examined students' and teachers' perceptions of automated scoring, and the students' and teachers' explanations of the quantitative results. Fourteen undergraduate students and nine faculty members had been interviewed, yielding three main themes: linguistic alienation, construct-irrelevant variance, and a need for human oversight.

Students had referred code switching as their natural resource for expressing complex ideas, especially when it comes to academic concepts, which are more understandable with the use of English terminology and Urdu explanation. Some participants had expressed the opinion that automated feedback only regarded mixed-language expressions as mistakes not taking into account meaning. These accounts helped to explain why having a lower score may impact students' confidence and trust in automated assessment. They also aligned with the researcher's own research on the psychological mechanisms of English language learning, and the concepts of cognitive load, anxiety, motivation, and self-efficacy (Mehmood et al., 2025). The current results did not confirm that AI scores led to persistent anxiety, but it did demonstrate that participants felt that the scoring process might be discouraging.

Teachers had indicated that assessment should be based on subject knowledge and not on accuracy of language and that automated scores should not be used to supplant teacher judgment. The concerns expressed by them were linked to the researcher's findings on teacher readiness for the introduction of AI and the evolving nature of teachers in the realm of AI education (Mehmood, Ahmad, & Jamil, 2026; Mehmood, Rubab, Iftikhar, & Goraya, 2026). These studies had already recognized the need for teacher preparation and institutional support and human oversight. The findings from the interviews therefore indicated that the success of the use of AI to assess a student's work required not only that the AI model performed well, but also that teachers are able and have the right to second guess the AI in making decisions.

5.3 Integration of Findings and Theoretical Implications

The quantitative and qualitative findings had converged around the central concern that code-switched writing was more difficult for the tested AI systems to score consistently. The quantitative analysis had identified larger AI–human differences and lower agreement for code-switched responses. Students and teachers had independently described experiences and concerns that were compatible with those patterns. This convergence strengthened the interpretation that Urdu-English code-switching needed explicit consideration in automated writing evaluation.

The findings also supported the study's combined theoretical perspective. Code-switching and translanguaging scholarship had helped explain why bilingual language use should be interpreted as a meaningful communicative practice rather than automatically treated as deficiency (García & Li Wei, 2014; Gumperz, 1982; Myers-Scotton, 1993). Construct-validity theory had clarified why an AI score could be unfair when it reflected language features outside the intended assessment construct (Kane, 2013). Algorithmic fairness had provided a basis for examining whether score differences were justified by the rubric. The researcher's work on dynamic bilingualism and pedagogical translanguaging had further highlighted the influence of institutional rules, teacher attitudes, and assessment expectations on the acceptance of bilingual practices.

5.4 Recommendations

Based on the findings, the following recommendations were developed:

1. Use construct-aligned rubrics. It is important that assessment rubrics clearly separate content, reasoning, organization and language proficiency. If the task allows, the use of bilingual expression should not lower the content scores.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.04 No.01 (2026)

2. Score in a bilingual fashion. Permission for Urdu-English code-switching, Roman Urdu variations and criteria to be evaluated should be explicitly communicated to the AI systems. Such prompts should be researched and not simply assumed to ensure a fair process.
3. Maintain human moderation. Scores produced by AI should be used as information for decision making. Disputed scores, unusually low scores, and scores that appear to have a great impact should be reviewed by the teacher or trained bilingual rater.
4. Audit systems in all languages. Institutions should be comparing the AI scores for each matched response (English only and code-switched responses), and tracking score differences, agreement, and consistency between models and tasks.
5. Provide teacher training. Professional development should include topics on AI literacy, rubric interpretation, linguistic diversity, and process for challenging and/or correcting AI scores. This recommendation aligns with the researcher's findings regarding teacher readiness for teaching with AI and institutional obstacles to inclusive AI use (Mehmood, Ahmad, & Jamil, 2026; Mehmood et al., 2026a).
6. Be transparent for students. Students should be notified of the use of AI, the reasons for using it and how to seek human input. Comments should be specific and point to aspects that could be improved rather than simply giving a numerical grade.
The limitations of the study are presented.The limitations of a study are presented.
The results were examined taking into account some limitations. First, the student writing samples used in the quantitative analysis were 120 samples and the students and faculty members in the qualitative phase were 14 and 9, respectively. All these participants and samples were not able to reflect the context of all Urdu-English bilingual learners and institutions in Pakistan. Second, the authors found that the language model, version, selected AI system, and selected prompts and writing tasks and rubrics were the factors they examined, and the findings may not be the same if these were not the same. Third, even though English and code-switched versions were matched with an attempt to keep the meaning same in both versions, the versions may not have been exactly the same in every linguistic detail. Fourth, human ratings had been established, but experience and interpretation could influence human judgment. Lastly, the regression and prompting results indicated relationships and changes across the tested design, but should not be replicated outside those conditions.

5.5 Limitations of the Study

It is suggested that larger and more diversified samples from various levels of education and different regions of Pakistan be used in future studies. They should replicate other AI models, record model versions, prompts, and scoring options, etc. There is need for further experiments, which must differentiate Urdu written in Perso-Arabic script from Roman Urdu and study the types and the frequencies of code-switching. The effectiveness of bilingual-aware prompts should also be examined with regard to writing genres, proficiency levels, and content areas. Future study might focus on the development of students' writing skills and confidence, and trust, as a result of multiple encounters with AI feedback. Finally, it is important to assess institutional moderation procedures and teacher training interventions in order to understand the translation of technical safeguards to fair assessment practice in the future.

5.7 Conclusion

This study had evaluated the fairness of AI-based writing assessment in Urdu-English bilingual students quantitatively by comparing the scores, and qualitatively through stakeholder interviews. The results had indicated that unprompted AI systems had resulted in greater differences in scores

and less agreement with human raters for code-switched writing, especially on a rubric that allowed for bilingual expression of the content. The reported disparity had been minimised with the use of bilingual-aware prompting, and interviews had brought up issues of linguistic alienation, construct validity concerns, and the need for human supervision. It is therefore, concluded that no pre-testing should be assumed to get the AI scorers to understand and evaluate code switching in Urdu-English in a fair manner. Construct-aligned rubrics, bilingual-aware directions, systematic audits, transparent feedback, and human meaningful review were essential for fair automated writing evaluation.

REFERENCES

Amjad, S., & Hussain, I. (2026). Roman Urdu and code-mixed language processing for social media analytics in Pakistan. Journal of Applied Linguistics and TESOL, 9(2). [Full publication details to be checked against the uploaded article.]

Choi, J., Tate, T., & Warschauer, M. (2026). Anchor is the key: Toward accessible automated essay scoring with large language model through prompting. Assessing Writing. https://doi.org/10.1016/j.asw.2026.101053

Debelak, R., & Ziegler, M. (2026). Testing standards for AI-based scores in automated essay scoring. PLOS ONE, 21(7), e0354680. https://doi.org/10.1371/journal.pone.0354680

Firoozi, T., Mohammadi, H., & Gierl, M. J. (2025). Using automated procedures to score educational essays written in three languages. Journal of Educational Measurement, 62(1), 33–56. https://doi.org/10.1111/jedm.12406

García, O., & Li Wei. (2014). Translanguaging: Language, bilingualism and education. Palgrave Macmillan.

Gumperz, J. J. (1982). Discourse strategies. Cambridge University Press.

Huang, Y., & Wilson, J. (2025). Evaluating LLM-based automated essay scoring: Accuracy, fairness, and validity. Proceedings of the Artificial Intelligence in Measurement and Education Conference: Works in Progress. https://aclanthology.org/2025.aimecon-wip.9/

Jiao, H., Song, D., & Lee, W.-C. (2026). Evaluating rater effects of large language models in automated essay scoring: GPT, Claude, Gemini, and DeepSeek. Educational Measurement: Issues and Practice. https://doi.org/10.1111/emip.70018

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73.

Mehmood, M. U., & Parveen, Z. (2021). Explore areas of needed support, competencies, attitudes, and concerns of elementary school teachers for inclusion of CWDS in elementary schools of Punjab. Global Educational Studies Review, 6(3), 109–120. https://doi.org/10.31703/gesr.2021(VI-III).12

Mehmood, M. U., & Parveen, Z. (2025). Empowering teachers for inclusive education: A support program for children with disabilities in Punjab's elementary schools. International Journal of Social Sciences Bulletin, 3(3), 395–405. https://doi.org/10.5281/zenodo.15010375

Mehmood, M. U., Jamal, B., Dara, D., & Goraya, M. M. (2025). The psychological processes behind English language learning and their role in achieving inclusive and equitable quality education. ACADEMIA International Journal for Social Sciences, 4(4), 6421–6439. https://doi.org/10.63056/academia.4.4.2025.1478

Mehmood, M. U., Zaman, Q.-u., Iftikhar, N., & Bhatti, A. U. R. (2025). Digital transformation in Pakistan's public schools: Opportunities and challenges for SDG-4. Journal of Applied Linguistics and TESOL, 8(3), 1669–1687.

Mehmood, M. U., Ahmad, S., & Jamil, B. (2026). Teachers' readiness to integrate artificial intelligence into inclusive classrooms. Journal of Applied Linguistics and TESOL, 9(1).

Mehmood, M. U., Khan, S., Rubab, U. E., Sama, A., & Goraya, M. M. (2026a). AI-powered curriculum strategies to promote inclusive and quality higher education. Qualitative Research Journal for Social Studies, 3(2), 450–475. https://doi.org/10.63878/qrjs1158

Mehmood, M. U., Rubab, U. E., Maqbool, S., & Goraya, M. M. (2026b). AI-generated content, digital misinformation, and critical media literacy among Pakistani university students: The roles of trust, cognitive reflection, and digital literacy. ACADEMIA International Journal for Social Sciences, 5(3), 1697–1722. https://doi.org/10.63056/academia.5.3.2026.2332

Mehmood, M. U., Rubab, U. E., Iftikhar, N., & Goraya, M. M. (2026c). Transforming the role of teachers in the age of AI: From knowledge providers to learning facilitators. Qualitative Research Journal for Social Studies, 3(1).

Mehmood, M. U., Parveen, Z., & Rubab, U. E. (2026). The action phase: Elaborating implementation and initial effects of a Teacher Support Program (TSP) for inclusive education in Punjab elementary schools. ACADEMIA International Journal for Social Sciences, 5(3), 1677–1696. https://doi.org/10.63056/academia.5.3.2026.2331

Myers-Scotton, C. (1993). Social motivations for codeswitching: Evidence from Africa. Oxford University Press.

Wu, H.-N., Chu, M.-N., & Hsu, J.-L. (2026). Comparing GPT and human raters in essay assessment: Variability, bias, and the potential of LLM-based scoring. Computers and Education Open, 10, 100341. https://doi.org/10.1016/j.caeo.2026.100341

Mehmood, M. U., & Parveen, Z. (2024a). A study to find out organizational stress among special education teachers in government special education centers in Punjab. Al-Qudwah, 2(4), 181–193.

Mehmood, M. U., Ain, Q. U., Jamal, B., & Bhatti, A. U. R. (2025). Integrating indigenous cultural practices into early childhood curricula: Challenges and opportunities. Journal of Political Stability Archive, 3(2), 480–499. https://doi.org/10.63468/jpsa.3.2.28

Mehmood, M. U., Bano, R., Zubair, A., & Bhatti, A. U. R. (2025). Psychological impact of languages on the human mind: Research on the contribution of psycholinguistics approach to teaching and learning English. Journal of Applied Linguistics and TESOL, 8(3), 1160–1174.

Mehmood, M. U., Fatima, A., Jamal, B., & Bhatti, A. U. R. (2025). The role of English linguistic competence in developing critical thinking, empathy, and psychological well-being through education. Contemporary Journal of Social Science Review, 3(4), 1616–1632.

Mehmood, M. U., Rehman, M., Sama, A., & Shaheen, B. (2025). The role of school leadership in promoting psychological safety and inclusion in public schools: Lessons for SDG-4 implementation. American Journal of Psychiatric Rehabilitation, 28(2), 185–195. https://doi.org/10.69980/ajpr.v28i2.840

Mehmood, M. U., Zaheer, K., Jamal, B., & Goraya, M. M. (2025). Motivation, language development, and educational success: A psychological and linguistic approach to

English teaching for sustainable learning outcomes (SDGs 4 & 8). Contemporary Journal of Social Science Review, 3(4), 1764–1781.

Mehmood, M. U., Ziauddin, & Naseem, A. (2025). Socio-cognitive barriers in adopting English as a medium of instruction in Pakistani secondary schools: Challenges and stakeholder perspectives. Al-Īmān Research Journal, 3(1), 122–137.

Mehmood, M. U., Rubab, U. E., Goraya, M. M., & Bhatti, A. U. R. (2026). First-language exclusion and socio-psychological well-being among marginalized children in early childhood education in Pakistan: An explanatory sequential mixed-methods study. Qualitative Research Journal for Social Studies, 3(1), 1279–1303.

Mehmood, M. U., Rehman, M., Sama, A., & Goraya, M. M. (2026). The evolving role of teachers and mentors in inclusive education: From facilitators to champions of diversity and sustainability in line with SDG 4 and SDG 5. ACADEMIA International Journal for Social Sciences, 5(3[s9]), 1–24. https://doi.org/10.63056/academia.5.3(s9).2026.2131

Mehmood, M. U., Jamil, B., Rubab, U. E., & Bhatti, A. U. R. (2026). Psycholinguistic interventions for enhancing social inclusion of non-verbal students through assistive technology. AL-AASAR Journal, 3(1), 545–563.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

Do AI Scorers Understand Urdu-English Code-Switching?

Fairness of Automated Writing Evaluation for Bilingual Learners

English? اردو؟

Both Together!

Bilal Jamil
M. Phil. English Linguistics
Emerson University Multan

Shahbaz Ahmad
M.Phil English Literature
University of Southern Punjab

Dr. Muhammad Umar Mehmood
PhD Education
Punjab School Education Department, Pakistan

1. Background

AI-based Automated Writing Evaluation (AWE) systems are widely used in education, but their fairness is uncertain, especially for **bilingual learners** who use **Urdu-English code-switching** in their writing.

Code-switching is a natural part of communication in Pakistan!



2. Research Aim

To examine whether AI-based scoring systems provide fair evaluation of **Urdu-English code-switched writing** among Pakistani undergraduate students, and how factors like bilingual-aware instructions and human supervision affect results.

3. Methodology

- **120 Writing Samples** (English-only & Code-Switched)
- **Two Rubrics:** Content & English-Proficiency
- **AI Systems & Human Raters**
- **14 Students & 9 Faculty** (Interviews)

4. Key Findings

1. AI scored Code-Switched writing much lower than Human (Mean Gap: −7.6 vs −0.9).
2. Agreement was lower for Code-Switched writing (ICC = 0.54 vs 0.82).
3. Code-Switching Density negatively correlated with AI scores ($R^2 = 0.38$, $p < .01$).
4. Bilingual-Aware Prompting reduced the score gap (from 7.6 to 1.4 points).

5. What Students & Teachers Say?



"Code-switching is part of our identity, not a mistake."
— Student



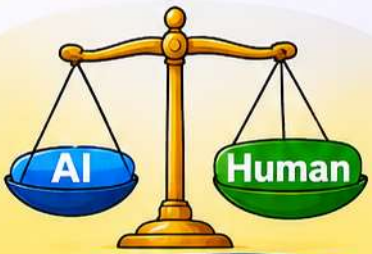
"AI should understand our mixed language context."
— Student



"Human supervision is essential for fair and valid scoring."
— Faculty Member

6. Implications

- Design rubrics aligned with multilingual learners' needs
- Use bilingual-aware instructions
- Conduct fairness audits
- Provide transparent feedback & train educators
- Ensure human moderation



Fair Assessment for Every Learner

7. Conclusion

AI systems should not automatically penalize bilingual expression when code-switching is allowed. With bilingual-aware design and human oversight, AI can support **equitable and valid** writing assessment in multilingual educational settings.

Let's build a fair, inclusive and multilingual future in education!