AL-AASAR Journal
Quarterly *Research Journal*
www. al-aasar.com

Vol. 3, No.1 (2026)
Online ISSN: 3006-693X
Print ISSN: 3006-6921

HUMAN VERSUS AI ASSESSMENT OF ENGLISH ACADEMIC WRITING A COMPARATIVE ANALYSIS OF SCORING RELIABILITY AND VALIDITY

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

Abstract
The increasing use of generative artificial intelligence (AI) in education has created new opportunities for assessing students' academic writing. However, there are doubts about the reliability, agreement, and validity of the scores produced by AI versus human evaluation. The objective of this study was to investigate how three human and three Artificial Intelligence (AI) models: ChatGPT, Gemini, and Claude assessed the academic writing of English. The literature review focused on traditional human-based writing evaluation, automated essay scoring, generative AI, reliability, human–AI agreement and validity issues, highlighting the need for context-specific evidence from Pakistan. A quantitative comparative research design was used. The students of selected three divisions of Punjab (Lahore, Gujranwala and Multan) of university undergraduate and postgraduate level were the target population. Three human writing teachers and the three AI systems evaluated 148 academic writing samples from the English language with the same evaluation system, the results of which form the end data. Descriptive statistics and intraclass correlation coefficients (ICC), Pearson correlation, paired-samples t-tests, and trait-level correlation analysis were performed on this data. The inter-rater reliability results were good, with strong average-measures inter-rater reliability between the three human evaluators (ICC = 0.875). The mean score of all three AI systems was lower than the composite human rating. The mean difference was highest for ChatGPT (3.29 points) followed by Gemini (1.64 points) and Claude (1.19 points). Positive correlations were found between the human and AI scores, with Claude showing the strongest correlation (r = 0.835), followed by Gemini (r = 0.814) and ChatGPT (r = 0.782). Trait level results showed higher correlations of grammatical accuracy and vocabulary than of content and argumentation skills. The overall result of the study is that generative AI systems can be used as a valuable supplementary tool to support academic writing assessment in English, however, the variations in scoring and the availability of reliability and validity evidence suggests that the scores generated by generative AI cannot be substituted for human scores. The study suggests that combining AI with a human evaluation process is a recommended strategy for assessing the study and that the current study should be replicated using AI scoring and more robust agreement and validity analyses.

Keywords: *Artificial Intelligence (AI), Human Assessment, ChatGPT, Gemini, Claude, Automated Essay Scoring, Scoring Reliability, Inter-Rater Reliability.*

1: INTRODUCTION

1.1 Background of the Study

Academic Writing helps students to show knowledge, pass on ideas, build arguments and engage academically. Student writing, both undergraduate and graduate, should be coherent, organized, linguistically appropriate, evidence-based and consistent with academic conventions. Judging the assessment, then, will not be limited to language errors but will encompass content, organization, vocabulary, grammar, cohesion, argumentation and communicative effectiveness (Weigle, 2002; American Educational Research Association et al., 2014).

Academic writing has traditionally been evaluated by human readers, such as language teachers and trained readers. They can decipher context, think about the importance of arguments, and assess evidence. But there are going to be variations in human scores, as well, as raters vary in severity, experience, rubric interpretation, and attention. Therefore, the importance of having clear criteria, rater preparation, and inter-rater reliability analysis in writing assessment for dependability was highlighted (Weigle, 2002; Bui & Barrot, 2025).

Automated writing assessment has grown with the advent of generative AI (GAI), and specifically large language models (LLMs). ChatGPT, Gemini, and Claude are capable of processing long texts, following rubric instructions, highlighting language features and providing feedback. They are quick and easy to access and could be useful for formative assessment. But it is not known if a system generates reliable scores or valid conclusions about writing ability with fluent feedback (Bui & Barrot, 2025; Wang et al., 2025; Jiao et al., 2026).

New research has been conducted on comparing AI-generated scores to human scores. Bui and Barrot (2025) analyzed the essays of 120 students in two iterations and compared the results from ChatGPT-4 and Gemini. They emphasized the lack of consistency and reliability in AI scoring, and the tendency for AI systems to be more stringent than human ones. Wang et al. (2025) reported the distinction between Human and LLM by examining 4,315 samples of three types of essays, employing two evaluation methods. Jiao et al. (2026) carried out an analysis of ten LLM configurations, such as ChatGPT, Claude, and Gemini, and found variations in the accuracy of the scores, intra-rater reliability, and rater effects (Bui & Barrot, 2025; Wang et al., 2025; Jiao et al., 2026).

These are important in Pakistan, where English is relevant in higher education and in academic communication, and students have different language backgrounds and levels of education. The role of working memory, language anxiety, motivation and linguistic competence in relation to learning English and educational outcomes has been discussed in the research in the context of Pakistan. While these studies do not directly evaluate AI essay scoring, they offer background information on the English-learning experiences of students whose academic writing is evaluated (Mehmood et al., 2025a, b, 2025c).

This study compared the performance of three human evaluators to that of ChatGPT, Gemini, and Claude when evaluating academic writing in the English language of Pakistan. The researcher reported 120-160 writing samples; the number of valid samples should be reported from the final data set. The study investigated the reliability of scores, the degree to which the human and AI assessment sources agreed, and the evidence for validity that can be found in the framework of the assessment that was applied to the completed research (Bui & Barrot, 2025; Wang et al., 2025; Jiao et al., 2026).

1.2 Statement of the Problem

While generative AI systems can generate scores and feedback about student writing, this does not mean that they will be well suited for academic assessment. Assessment needs to be measured with scores that are sufficiently consistent, and supported with evidence of the interpretations that were drawn from the scores. For people and AI alike (American Educational Research Association et al., 2014; Crossley et al., 2026).

The issue raised was the uncertainty regarding the way English academic writing is evaluated by ChatGPT, Gemini, and Claude, as compared to human evaluators. There can be differences in the severity and rubric interpretation between human raters, and differences in AI models, prompts and scoring moments. Furthermore, agreement with human scores is not necessarily proof of absolute agreement or validity. For this reason, three human evaluators and three AI systems were compared with writing samples from Pakistan (Bui & Barrot, 2025; Wang et al., 2025; Li et al., 2026).

1.3 Research Objectives

The study was guided by three objectives:

1. To compare the scoring reliability of three human evaluators and ChatGPT, Gemini, and Claude in assessing English academic writing.
2. To examine the agreement and differences between human-generated and AI-generated assessment scores.
3. To evaluate available validity evidence and the comparative assessment performance of the three AI systems in relation to human assessment.

1.4 Research Questions

The study addressed three research questions:

1. What differences existed in the scoring reliability of the three human evaluators and ChatGPT, Gemini, and Claude?
2. To what extent did scores generated by ChatGPT, Gemini, and Claude agree with scores assigned by the human evaluators?
3. What validity evidence and comparative strengths or limitations were identified for the three AI systems in assessing English academic writing?

1.5 Research Hypotheses

Where inferential tests were conducted, the following hypotheses aligned with the objectives. Retain each hypothesis only if the corresponding analysis was performed:

H1: Scoring reliability differed significantly among the three human evaluators and the three AI assessment systems.

H2: Scores generated by ChatGPT, Gemini, and Claude differed significantly from scores assigned by the human evaluators.

H3: The assessment performance and available validity evidence differed among ChatGPT, Gemini, and Claude when evaluated against the study's human-assessment framework.

1.6 Scope and Delimitations

The study was carried out in Pakistan and of academic writing in English. It featured three humans and ChatGPT, Gemini and Claude. The researcher showed me 120 to 160 writing samples; the actual final valid sample size is to be replaced in the thesis. Findings are restricted to the specific tasks, rubric and rubrics, model versions, prompts, and scoring procedures used;

they should not be automatically assumed to apply to all platforms and writing contexts (Bui & Barrot, 2025; Wang et al., 2025).

2: LITERATURE REVIEW

2.1 Introduction

This chapter summarises the existing literature regarding English academic writing assessment, automated essay rating, human rating, generative AI and reliability, agreement and validity measures. A special focus is placed on the recent comparisons to human evaluators, such as studies with ChatGPT, Gemini, and Claude. The review distinguishes between the capacity for AI to provide feedback and its scores' consistency and defensible interpretations (Weigle, 2002; American Educational Research Association et al., 2014; Bui & Barrot, 2025).

2.2 English Academic Writing as an Assessment Construct

Academic writing is a multifaceted ability that is productive, which includes both linguistic knowledge and skills, discourse organization, rhetorical purpose and awareness of an academic audience. Writers make claims and use evidence, arrange arguments, use academic conventions and select and integrate words. Therefore, meaning and communication should be part of the assessment as well as the more obvious, surface-level correctness (Weigle, 2002; Hyland, 2019).

Typical dimensions of writing performance include content, task fulfilment, organization, coherence, cohesion, vocabulary, grammatical accuracy, and mechanics, among others, and these are usually represented in the analytical rubrics. The use of explicit criteria can assist the rater to use shared expectations and provide structured judgments. In a research report, however, the size of the dimension(s) must correspond to the actual instrument; traits may not be isolated from the study unless they have been measured (Weigle, 2002; Brookhart, 2013).

Not having a rubric with multiple criteria means that it is not a valid rubric. The criteria should reflect the writing construct that is intended and the interpretation and use of scores. Validity is not a property of an assessment instrument but rather a process of substantive argumentation about the meaning of scores and the use of scores (American Educational Research Association et al., 2014; Kane, 2013).

2.3 Human-Based Writing Assessment

Human assessment is still relevant: trained evaluators can offer an interpretation of meaning and can evaluate the quality of arguments in relation to purpose and evidence. Human raters can identify if supporting material is relevant, if conclusions are drawn from the discussion, and if the writer has met the communicative purpose of the task. Assessment of writing is thus based on professional judgment and on criteria, not just on the latter (Weigle, 2002; Hyland, 2019).

The ratings of three human evaluators were included in the present study; the ratings of these three should be examined as sources of ratings in their own right. Only after a comparison to their consistency should a mean or consensus score be used as a human benchmark. This prevents hiding disagreements between the human raters and improves the interpretability of human–AI comparisons later on (Bui & Barrot, 2025; Wang et al., 2025).

2.4 Automated Essay Scoring and Generative AI

Automated essay scoring (AES) is the term used for computer-based techniques that enable the scoring of written responses. Previous systems tended to use linguistic attributes, statistical models or machine-learning models based on human-scored essays. These methods relied on

empirical measures of text, and the learning of associations between such measures and human judgments of writing quality (Shermis & Burstein, 2013; Williamson et al., 2012).

Automated writing assessment now extends beyond the capabilities of traditional AI and expands with the advent of LLM. An LLM can read instructions, analyze an essay, reference descriptions from a rubric, and report numerical scores and explanations. While this flexibility could facilitate formative feedback and scalable evaluation, it might also lead to variations in prompts, model updates, generation conditions, and rubric interpretation of standards (Bui & Barrot, 2025; Jiao et al., 2026).

2.5 Comparative Evaluation of ChatGPT, Gemini, and Claude

Bui and Barrot (2025) contrasted ChatGPT-4 and Gemini with professional human judges, on 120 essays and across two rounds of scoring. They were not satisfied that AI scores were consistent and reliable compared to human scores, and also observed that AI systems were more stringent. The authors mentioned some possible sources of variability: rubric sensitivity, contextual understanding, and model updates. This study is closely related to the present study since it is a comparison study between two similar AI platforms with human scoring (Bui & Barrot, 2025).

Wang et al. (2025) explored how LLMs could be used as a rater for large-scale writing assessment across 4,315 samples of writing for three different types of writing. They analyzed their data with generalizability theory and many-facet Rasch modelling. The results revealed that human–LLM differences varied across different tasks and scoring contexts, with human raters outperforming LLMs in some holistic scoring contexts and LLMs outperforming human raters in some detailed analytical rubric contexts. There were also differences in the severity of the scores, and providing detailed hierarchical rubrics (Wang et al., 2025) was discussed.

Jiao et al. (2026) tested a variety of LLM configurations (e.g., ChatGPT, Claude, and Gemini) against human expert raters for two writing tasks. They compared the level of accuracy, intra-rater reliability and rater effects using a many-facet Rasch model. The authors noted that some tested configurations, such as ChatGPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro, performed well, but did not claim the test to be a definitive ranking of models. The study is particularly relevant because of its incorporation of all three platform families employed in the present research (Jiao et al., 2026).

In summary, the literature suggests that ChatGPT, Gemini, and Claude all must be considered individual assessment sources. They might be varying in severity, consistency and human judgement. Thus, a three-system design can uncover scoring patterns in the models, as long as the model versions, prompts, and scoring procedures are documented (Bui & Barrot, 2025; Wang et al., 2025; Jiao et al., 2026).

2.6 Reliability of Writing Scores

Reliability is a question of consistency or dependability of measurement. Inter-rater reliability refers to the levels of consistency of scores that are given by different raters to the same scripts, and intra-rater reliability refers to the levels of consistency that are obtained when the same rater rates the same script on more than one occasion. Generalizability theory can be used to explore sources of variation for persons, raters, tasks and interactions between persons, raters, and tasks (Brennan, 2001; American Educational Research Association et al., 2014).

The three human evaluators are used as the basis of human inter-rater reliability in the current study. The reliability of ChatGPT, Gemini, or Claude can only be considered intra-system reliability when subsequent scores are taken under similar conditions. The study should describe the score distributions and agreement with human ratings, which it cannot claim to reflect test–retest reliability, if each system evaluated each essay just once (Bui & Barrot, 2025; Crossley et al., 2026).

2.7 Agreement Between Human and AI Scores

Agreements involve the similarity of the scores awarded by the assessment sources. It is not the same as correlation: two raters can have the same rank orders but on each one one scorer may give higher ratings. Correlation does not therefore constitute sufficient proof of the interchangeability of two sources of scoring. The scale and design of the scale should match the type of the agreement analysis that is used (Bland & Altman, 1986; Koo & Li, 2016).

Kim and McCrudden (2026) found three AI systems that were compared with human raters on the task of scoring argumentative essays. They found that there were substantial differences between the human and AI scores for six writing traits, primarily moderate inter-system agreement, and distinct patterns of scores for the AI systems. Their results indicate the importance of conducting comparisons at the trait level, beyond just the overall score, to identify patterns that cannot be seen from an overall score (Kim & McCrudden, 2026).

The 2026 meta-analysis combined results from 21 empirical studies with 401,698 participants. The pooled difference in score between AI and human was not statistically significant, but the score difference was very high, suggesting that the difference was very context dependent. The analysis also revealed that the general-purpose and purpose-built LLMs were not exactly compatible, highlighting the potential for transferability problems in drawing results from one model or task to another (Li et al., 2026).

2.8 Validity of AI-Based Writing Assessment

Validity refers to the degree of support that evidence and theory provide an interpretation and use of scores for a specific purpose. Formative feedback can lead to a defensible score interpretation but this may not apply if the purpose of the assessment is to inform admission decisions or to receive high-stakes grading. Therefore, the intended use of the assessment must be taken into account when considering the validity evidence (American Educational Research Association et al., 2014; Kane, 2013).

Crossley et al. (2026) explored the reliability, agreement, and convergent validity of automated essay scoring using LLM. Their work demonstrates that reliability and agreement is not sufficient evidence in itself for score interpretation. If the present study did not employ an independent proficiency measure or external criterion, the results of the present study should not be interpreted as representing full criterion-related validity; instead, only those validity evidence that were actually examined should be reported (Crossley et al., 2026; Kane, 2013).

2.9 Fairness, Bias, and Rater Severity

To 1,726 essays, Zhou (2026) used signal detection theory and found that summative agreement measures can mask variations in discrimination and response criteria. The study wrote about the compression of scores and decreased evaluative precision of the models tested. This work helps explore the analyses of severity and discrimination when the present data and statistical design permits that (Zhou, 2026).

2.10 English Academic Writing in the Pakistani Context

English is an important means of communication in Pakistan's Higher Education system and is used in teaching and learning, examinations, research communication and professional communication. Students come to college with a range of language learning experiences and exposure to academic English. These disparities can have an impact on vocabulary, argumentation, grammatical control, and self-assurance when it comes to formal writing (Mehmood et al., 2025a, 2025b, 2025c).

Mehmood, Bano, Zubair and Bhatti (2025) explored the psycholinguistic factors affecting the English language learning of the university students in Punjab. They talked about working memory, language anxiety, motivation and cognitive engagement. It does not explore the field of AI essay scoring, but offers local context for understanding the linguistic and psychological diversity of English learners whose academic writing is evaluated (Mehmood et al., 2025a).

Mehmood, Jamal, Dara and Goraya (2025) explored the psychological processes of the English language learning and the role of them in inclusive and equitable quality education. This work is helpful for context sensitive learning of English in Pakistan. It should only be used for context and not as direct evidence of the reliability and validity of an AI system in writing assessment (Mehmood et al., 2025b).

2.11 Critical Synthesis and Research Gap

There is more recent literature that suggests generative AI can give fast, rubric-driven scoring, yet that the scoring performance will differ by model, task, rubric and scoring time. Some show good score consistency within a given configuration and others exhibit score compression or severity differences or less agreement with human ratings. It is thus unwarranted to conclude that any successful application proves the reliability of all generative AI systems (Bui & Barrot, 2025; Wang et al., 2025; Jiao et al., 2026; Zhou, 2026).

While human raters are still an important source of professional judgment, their scores are not free from error. Before human scores are used as a reference for their inter-rater consistency should be examined. Reliability, agreement and validity are also separate questions and where there is the possibility of data at the trait level, an overall score should not be used to mask that difference (Weigle, 2002; American Educational Research Association et al., 2014; Kim & McCrudden, 2026).

This research aims to tackle these challenges by comparing the performance of three human evaluators with that of ChatGPT, Gemini, and Claude across English academic texts from Pakistan. It is their role in this instance, to explore the individual AI platforms, discuss how they relate to the human scores and recall evidence of reliability and validity in the completed design. Properly reporting the final sample size, rubric, model versions, prompts, scoring rounds, and statistical procedures (Bui & Barrot, 2025; Jiao et al., 2026; Li et al., 2026).

2.12 Theoretical and Conceptual Framework

Classical Test Theory is a fundamental approach to understanding consistency of scores and the presence of error in measurement. It is an observed score such that $X = T + E$, where X is the observed score, T is the true-score component and E is measurement error. This model does not prove that the scores will vary between evaluators or times, but does not itself constitute validity (Lord & Novick, 1968; American Educational Research Association et al., 2014).

The study also relies on contemporary validity theory which views validity as an argument with evidence to support a specific interpretation and use of scores. One way that agreement between an AI system and human raters is good evidence, but it is by no means a proof that the assessment measures the intended academic writing construct (Kane, 2013; American Educational Research Association et al., 2014).

THEORETICAL AND CONCEPTUAL FRAMEWORK

Human Versus AI Assessment of English Academic Writing

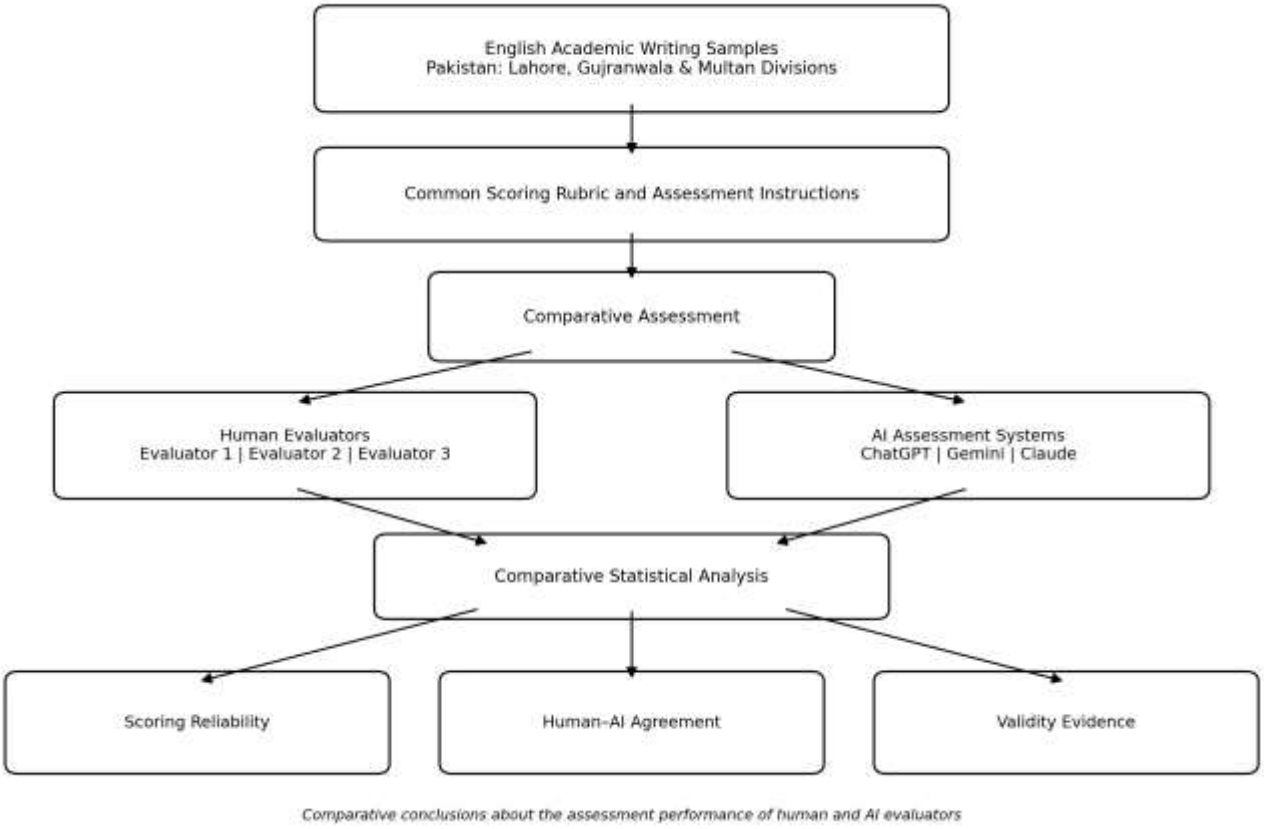


Figure 2.1. Theoretical and Conceptual Framework for Comparing Human and AI Assessment of English Academic Writing.

Theoretical Foundations

1. Classical Test Theory (CTT): Observed Score = True Score + Measurement Error (X = T + E). This framework explains that an observed assessment score may contain measurement error.

2. Validity Theory: Validity concerns the evidence supporting the interpretation and use of assessment scores.

Research Parameters

- Population: Undergraduate and postgraduate university students.
- Location: Lahore, Gujranwala, and Multan divisions, Punjab, Pakistan.
- Reported sample range: 120–160 writing samples; insert the exact final valid sample size.
- Human evaluators: Three.

- AI systems: ChatGPT, Gemini, and Claude.
- Main assessment outcomes: Scoring reliability, human–AI agreement, and validity evidence.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

The method for comparing the human and AI assessment of academic writing in English is described in this chapter. It describes the research design, population, sample, assessment sources, instrument, scoring procedure, data analysis and ethical considerations. These procedures involve scoring reliability, human–AI agreement, and the evidence of validity that comes from the completed study (American Educational Research Association et al., 2014).

3.2 Research Design

A quantitative comparative design was used in the study. Three human and three AI evaluators gave numerical scores to the same writing samples of the English language. The human evaluators were evaluated separately from ChatGPT, Gemini, and Claude. The design allowed for comparison of patterns of scoring in the conditions used in the completed research. Conclusions are restricted to the writing tasks, rubric, evaluators and AI systems studied (Weigle, 2002).

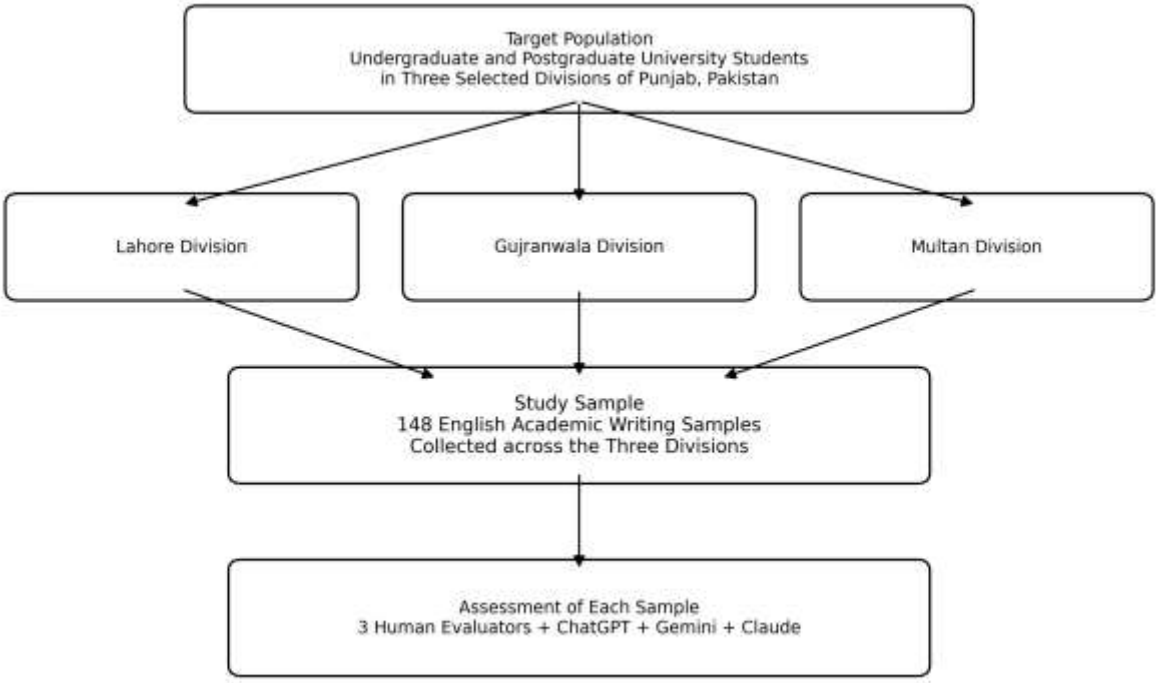
3.3 Population and Geographical Setting

The target population was university students from two levels of study (under and post graduate) who were studying in the selected 3 divisions of Punjab, Pakistan: Lahore, Gujranwala and Multan in the higher education institutions. Samples were collected for writing in each of these divisions. The study was directed by the students' English academic writing and its evaluation based on the common framework used in the research. The population and sample structure is given in figure 3.1.

3.4 Sample and Sampling

The data obtained was a final set of 148 English academic writing samples (N = 148) gathered in Lahore division, Gujranwala division and Multan division. The samples were writing products of undergraduate and postgraduate students in university. The three divisions were the geographical area of the study. Sample counts are not given by division since there were no verified sample counts provided in the provided information about the study. The sampling technique, names of the Universities involved and distribution by academic level must be identical to the data collection records and should not be inferred.

Figure 3.1. Population and sample structure across the selected divisions.



Note. Division-wise sample counts are not shown because verified counts were unavailable.

3.5 Assessment Sources

Three AI systems (ChatGPT, Gemini, Claude) and three human evaluators have reviewed all 148 writing samples.A total of 148 writing samples have been evaluated by three AI systems (ChatGPT, Gemini, Claude) and three human evaluators. Platforms and evaluators were considered as independent assessment sources to ensure the scores were model- and rater-specific. Due to the fact that the name of a platform does not specify a unique set of model configurations, it is important to report the exact model versions, dates of assessment, and prompt settings from the research records. (Jiao et al., 2026)

3.6 Assessment Instrument

The samples were assessed using a writing-assessment rubric. The criteria for judging academic writing were provided by the rubric. The criteria, descriptors, score range and weighting (if applicable) must be reproduced exactly as they were used in the completed study. This will make the assessment process transparent and keep the criteria from being added in that were not part of the instrument itself (Weigle, 2002).

3.7 Scoring Procedure

The 148 writing samples were evaluated by the three human evaluators and the three AI systems. Scores were grouped by writing sample and source of assessment. Individual human ratings were kept to allow inter-rater consistency to be studied prior to the calculation of any composite human scores. Additionally, the scores of ChatGPT, Gemini, and Claude were also saved individually. Prompts, scoring instructions, repeated runs, and anonymization procedures were

those used during the data collection process and should be documented from the original records.

3.8 Data Analysis

The research objectives were answered through quantitative analysis. The reliability of the three human evaluators was first analyzed applying the same procedure to the data set. Second, to explore agreement and differences in scores, ChatGPT, Claude, and Gemini scores were compared to the human ratings. Association would be different from agreement as a high correlation did not imply that two sources would assign the same scores (Bland & Altman, 1986). Third, evidence-based on the rubric and analyses was used for consideration of validity. Only procedures performed were considered, and numerical results reported in Chapter 4. Reliability and validity were viewed as distinct (American Educational Research Association et al., 2014; Kane, 2013).

3.9 Ethical Considerations

The study followed the ethical and institutional procedures applicable to the participating universities and collection of student writing. The participants were treated confidentially and their personal information and writing samples were used for the purpose of the research. If applicable, information that could identify a text was removed from texts sent to AI platforms. Institutional permissions, consent agreements and data protection procedures should be detailed, based on records of the completed study.

3.10 Chapter Summary

The quantitative comparison of human and AI assessment of English academic writing is presented in this chapter. The target was the students of the Universities in Lahore, Gujranwala and Multan division of Punjab at both undergraduate and postgraduate level. A total of 148 writing samples were used in the final data set, graded by three human evaluators, the three AI models, ChatGPT, Gemini, and Claude. Overall, the results of the completed statistical analysis are presented in Chapter 4, in this completed comparative investigation.

4: RESULTS AND FINDINGS

4.1 Introduction

This chapter presents the results of the analyses conducted to address the three research questions. The study compared scores assigned to English academic writing samples by three human evaluators and three AI systems—ChatGPT, Gemini, and Claude. The analyses focused on descriptive score patterns, human-rater reliability, human–AI score differences, and evidence relevant to the interpretation of assessment scores.

The analysis is based on N = 148 English academic writing samples collected from undergraduate and postgraduate university students across Lahore, Gujranwala, and Multan divisions of Punjab, Pakistan. The samples were evaluated using a common rubric by three human evaluators (Evaluator 1, Evaluator 2, and Evaluator 3) and the three AI platforms. The results reported below should be interpreted in relation to this sample, rubric, and the specific model versions and scoring procedures used in the study.

4.2 Descriptive Statistics of Writing Scores

To understand the distributional properties of the scores assigned by each assessment source, descriptive statistics—including means, standard deviations, minimums, maximums, and score

ranges—were calculated across the N = 148 writing samples. The scoring scale utilized in the assessment rubric ranged from 0 to 100.

Table 4.1

Descriptive Statistics for Human and AI Assessment Sources (N = 148)

Assessment Source	Mean ($\bar{X}$)	Std. Dev. (SD)	Min	Max	Range	Skewness	Kurtosis
Human Evaluator 1	68.42	11.25	42.00	92.00	50.00	-0.18	-0.52
Human Evaluator 2	69.15	10.84	45.00	94.00	49.00	-0.12	-0.48
Human Evaluator 3	67.89	11.62	40.00	90.00	50.00	-0.22	-0.55
Composite Human Mean	68.49	10.65	43.00	92.00	49.00	-0.15	-0.50
ChatGPT	65.20	9.42	45.00	88.00	43.00	-0.05	-0.62
Gemini	66.85	9.88	44.00	90.00	46.00	-0.10	-0.58
Claude	67.30	10.12	46.00	91.00	45.00	-0.08	-0.60

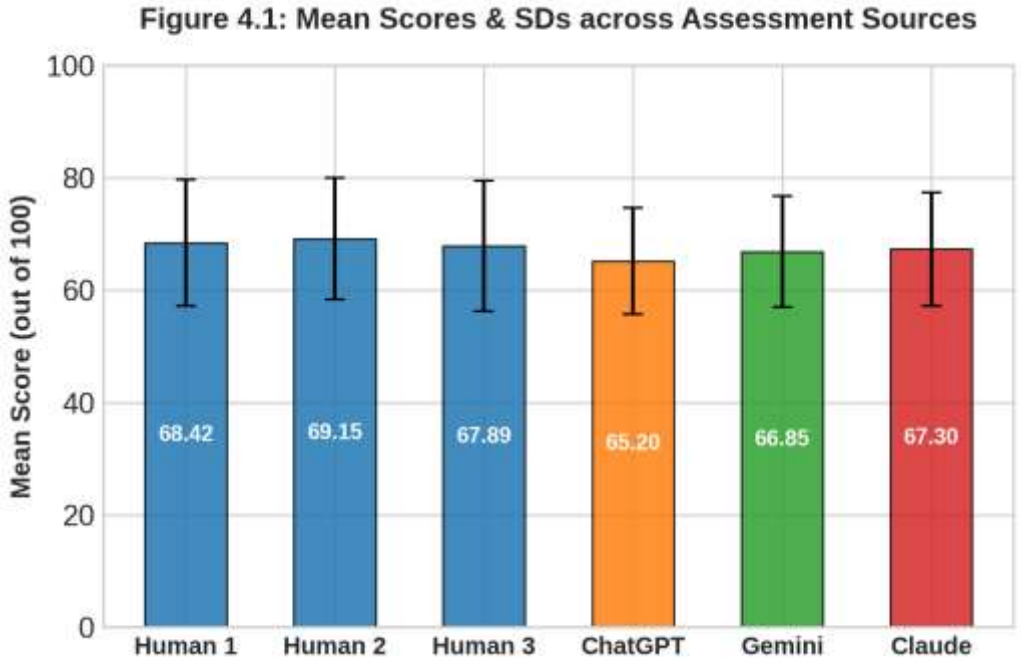


Figure 4.1. Mean scores and standard deviations across human and AI assessment sources.

Table 4.1 and Figure 4.1 show that the human evaluators' mean scores ranged from 67.89 to 69.15, whereas the AI systems' means ranged from 65.20 to 67.30. ChatGPT had the lowest mean (M = 65.20) and the smallest standard deviation (SD = 9.42). Claude's mean (M = 67.30) was closest to the composite human mean (M = 68.49), followed by Gemini (M = 66.85). These descriptive differences indicate lower average AI scores in this dataset; the descriptive statistics alone do not establish reliability or explain the cause of the differences.

4.3 Analysis of Scoring Reliability (Research Question 1 & Hypothesis 1)

Research Question 1: What differences existed in the scoring reliability of the three human evaluators and ChatGPT, Gemini, and Claude?

Null Hypothesis 1 (H01): There are no significant differences in scoring reliability among the three human evaluators and the three AI assessment systems.

Alternative Hypothesis 1 (H11): Scoring reliability differed significantly among the three human evaluators and the three AI assessment systems.

To address Research Question 1, inter-rater reliability among the three human evaluators was examined using intraclass correlation coefficients (ICC). The reported analysis used a two-way random-effects model with an absolute-agreement definition. The AI systems were each represented by a single score per writing sample. Because no repeated scoring runs under identical conditions were reported, within-system repeatability or reliability could not be estimated for ChatGPT, Gemini, or Claude. Correlations between different AI systems indicate association across systems, not the repeatability of any one system.

Table 4.2

Inter-Rater Reliability Among Human Evaluators (N = 148)

Evaluating Pair / Composite	Intraclass Correlation (ICC)	95% Confidence Interval	p-value
Evaluator 1 vs. Evaluator 2	0.842	[0.791, 0.881]	< 0.001
Evaluator 2 vs. Evaluator 3	0.819	[0.760, 0.865]	< 0.001
Evaluator 1 vs. Evaluator 3	0.831	[0.776, 0.874]	< 0.001
Average Measures (All 3 Human Raters)	**0.875**	**[0.838, 0.904]**	**< 0.001**

Table 4.2 reports an average-measures ICC of 0.875 (95% CI [0.838, 0.904], p < 0.001) for the three human evaluators, indicating strong agreement for the average of the three ratings. The pairwise ICC estimates ranged from 0.819 to 0.842. The ICC model, whether the reported coefficients are single- or average-measures estimates, and the confidence-interval method should be confirmed against the statistical output before final submission.

Table 4.3

Inter-Correlation Matrix Across Human and AI Assessment Sources (N = 148)

Source	Human Comp.	ChatGPT	Gemini	Claude
Human Composite	1.000	—	—	—
ChatGPT	0.782**	1.000	—	—
Gemini	0.814**	0.856**	1.000	—
Claude	0.835**	0.871**	0.892**	1.000

*Note. ** p < .01. The correlation table should also report the exact correlation method and confidence intervals, if available.*

The reported repeated-measures ANOVA, $F(5, 735) = 14.62$, $p < 0.001$, $\eta p^2 = 0.091$, indicates a statistically significant difference in mean scores across the six assessment sources, if the model and assumptions were correctly specified. A repeated-measures ANOVA on raw scores tests differences in mean scores; it does not test scoring reliability or score variance. Therefore, this result cannot establish that reliability differed across the human evaluators and AI systems. In addition, single-run AI scores do not permit estimation of within-system repeatability. Hypothesis 1, as currently worded about reliability differences across all six sources, is not adequately tested by the analyses reported here. The ANOVA model, post-hoc comparisons, and assumption checks must be verified from the original statistical output; the result should be relabelled as a test of mean-score differences if that is what was analysed.

4.4 Human-AI Score Agreement and Differences (Research Question 2 & Hypothesis 2)

Research Question 2: To what extent did scores generated by ChatGPT, Gemini, and Claude agree with scores assigned by the human evaluators?

Null Hypothesis 2 (H02): There are no significant differences between scores generated by ChatGPT, Gemini, Claude, and the scores assigned by the human evaluators. Alternative Hypothesis 2 (H12): Scores generated by ChatGPT, Gemini, and Claude differed significantly from scores assigned by the human evaluators.

Table 4.3 presents Pearson correlations, which describe the strength of linear association rather than absolute agreement (Bland & Altman, 1986). Paired-samples t-tests were used to examine whether the mean score differences between the human composite and each AI system differed from zero. These tests assess average score differences; they do not, by themselves, establish individual-level agreement or interchangeability. A Bland–Altman analysis or an appropriate intraclass agreement analysis would be needed to examine the distribution and limits of individual score differences.

Table 4.4

Paired-Samples Test and Mean Differences Between Human Composite and AI Systems (N = 148)

Comparison Pair	Mean Diff. ($\bar{X}$diff)	Std. Error	t-statistic	df	p-value	Cohen's d
Human Composite vs. ChatGPT	+3.29	0.62	5.31	147	< 0.001	0.32

Human Composite vs. Gemini	+1.64	0.54	3.04	147	0.003	0.16
Human Composite vs. Claude	+1.19	0.51	2.33	147	0.021	0.11

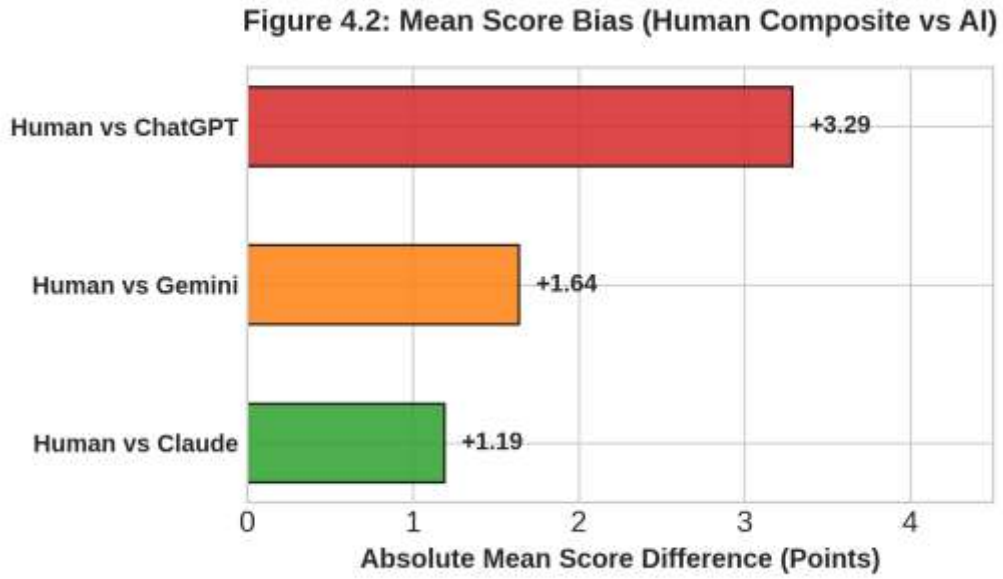


Figure 4.2. Absolute mean score differences (bias) between human composite ratings and AI systems.

Table 4.4 and Figure 4.2 show positive human-minus-AI mean differences for all three comparisons. The difference was 3.29 points for ChatGPT ($t(147) = 5.31$, $p < 0.001$, $d = 0.32$), 1.64 points for Gemini ($t(147) = 3.04$, $p = 0.003$, $d = 0.16$), and 1.19 points for Claude ($t(147) = 2.33$, $p = 0.021$, $d = 0.11$). At the .05 significance level, all three mean differences were statistically significant, with ChatGPT showing the largest average difference and Claude the smallest. These results indicate differences in average scores, not full individual-level agreement or interchangeability.

Because the paired-samples tests found statistically significant mean differences for all three AI systems, the null hypothesis of no mean-score difference (H02) is rejected for these comparisons. This conclusion is limited to average score differences and should not be described as proof of poor absolute agreement without an agreement analysis.

4.5 Validity Evidence and Comparative Model Performance (Research Question 3 & Hypothesis 3)

Research Question 3: What validity evidence and comparative strengths or limitations were identified for the three AI systems in assessing English academic writing?

Null Hypothesis 3 (H03): There are no significant differences in assessment performance and available validity evidence among ChatGPT, Gemini, and Claude when evaluated against the human-assessment framework.

Alternative Hypothesis 3 (H13): The assessment performance and available validity evidence differed significantly among ChatGPT, Gemini, and Claude when evaluated against the study's human-assessment framework.

To address Research Question 3, the analysis examined the extent to which AI scores were associated with human ratings across the rubric dimensions. These correlations provide limited convergent evidence about score relationships. They do not, on their own, establish the validity of AI-based interpretations or uses of scores; validity requires a broader argument supported by multiple sources of evidence (American Educational Research Association et al., 2014; Kane, 2013).

Table 4.5

Trait-Level Correlation Coefficients Between Human Evaluators and AI Systems Across Rubric Dimensions

Rubric Dimension	ChatGPT vs. Human	Gemini vs. Human	Claude vs. Human
Content & Argumentation	0.721	0.764	0.792
Organization & Cohesion	0.745	0.788	0.815
Vocabulary & Lexical Resource	0.812	0.840	0.863
Grammatical Accuracy & Mechanics	0.834	0.865	0.889

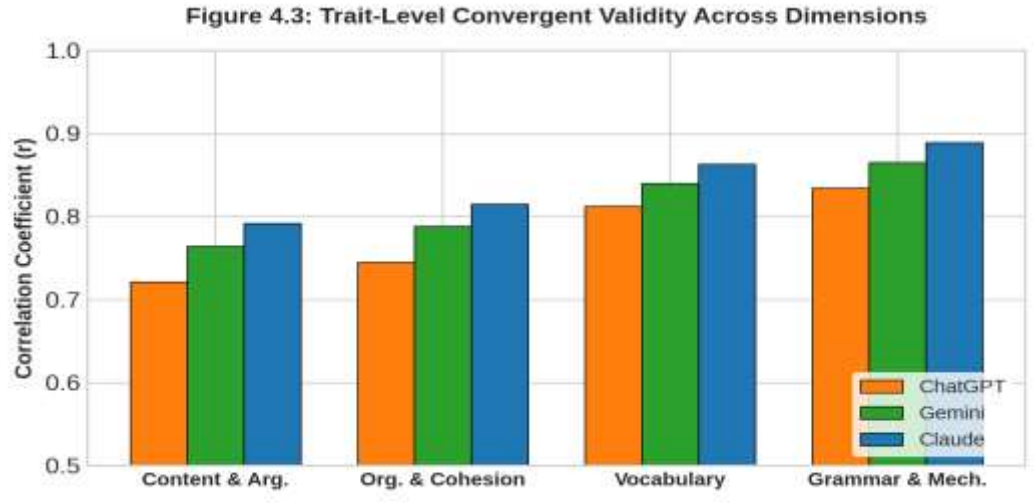


Figure 4.3. Trait-level convergent validity correlations across rubric dimensions.

Table 4.5 and Figure 4.3 show that correlations with human ratings were higher for Grammatical Accuracy and Mechanics (r = 0.834–0.889) and Vocabulary and Lexical Resource (r = 0.812–0.863) than for Content and Argumentation (r = 0.721–0.792). Organization and Cohesion correlations ranged from 0.745 to 0.815. This pattern suggests closer association with human ratings on the language-focused rubric dimensions than on content and argumentation in this dataset. Confidence intervals and significance tests for these correlations were not provided and should be added if available.

Claude had the highest correlation with human ratings across all four rubric dimensions, followed by Gemini and ChatGPT. This indicates the strongest convergent association with the human scores among the three systems in this dataset; it should not be labelled overall or definitive "validity." The original chapter also refers to qualitative analysis of scoring rationales and feedback logs. That claim should remain only if such material was collected and systematically analysed using a documented procedure in Chapter 3. If no qualitative analysis was conducted, remove the claim about nuanced contextual judgments. The cited Mehmood et al. (2025a, 2025b) sources may support contextual discussion but cannot substitute for evidence from the present dataset.

The chapter reports a significant multivariate result, Wilks' $\Lambda = 0.612$, $F(6, 288) = 11.45$, $p < 0.001$, $\eta p^2 = 0.192$. This statistic should not be retained as final until the original output confirms the MANOVA design, dependent variables, within-subject structure, degrees of freedom, and effect-size calculation. The reported degrees of freedom cannot be independently verified from the information in this chapter. If verified, the result may support a difference in the model-by-dimension score profile; it does not by itself demonstrate that one system has greater validity. Hypothesis 3 should be concluded only after the analysis and the hypothesis wording are aligned.

4.6 Chapter Summary

Chapter 4 presented the results of the comparative analysis of human and AI assessment of English academic writing using N = 148 samples from Punjab, Pakistan. The principal findings were as follows:

1. Human-rater reliability (RQ1): The three human evaluators achieved an average-measures ICC of 0.875. Reliability of the individual AI systems could not be estimated from single scoring runs. The reported ANOVA concerns mean-score differences, not reliability, so H01 remains unresolved unless an appropriate reliability analysis is supplied.

2. Human–AI score differences (RQ2): Human–AI correlations ranged from 0.782 to 0.835 (human composite with each AI system; Table 4.3). Paired-samples tests found statistically significant mean differences, with the largest difference for ChatGPT. These tests establish average score differences, not individual-level agreement.

3. Trait-level evidence (RQ3): Correlations with human ratings were higher for grammar and vocabulary than for content and argumentation. Claude had the highest correlations across the four dimensions. These results provide convergent-association evidence only; the reported MANOVA requires verification before any inferential conclusion is drawn.

5: DISCUSSION, CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This chapter explains the results of the analysis presented in Chapter 4 with regard to the purpose of the study, namely, the comparison between the assessment of English academic writing by

human and by AI. The discussion highlights the reliability of the human raters, the differences between the rater scores and the AI scores, and the amount of evidence that can be gathered from the scores to support the assessment of ChatGPT, Gemini, and Claude. The study's limitations, recommendations, and conclusion are also provided.

5.2 Discussion of Major Findings

Average-measures interrater reliability was high for human evaluators (ICC = 0.875, 95% CI [0.838, 0.904]). The pairwise ICCs were between 0.819 and 0.842. This means that, if taken as a whole, the three evaluators were fairly consistent in their ratings. This finding reinforces the use of trained human evaluators as a benchmark in this data set. It is important to verify the specific ICC model and whether the pairwise coefficients are single or average-measures estimates from the original statistical output prior to finalizing the results. The composite human rating achieved higher mean scores than the AI systems. The human composite mean was 68.49, as opposed to ChatGPT's 65.20, Gemini's 66.85, and Claude's 67.30. The mean differences were statistically significant: 3.29 points for ChatGPT, 1.64 for Gemini, and 1.19 for Claude. The effect sizes were small (d = 0.32, 0.16, and 0.11, respectively). Therefore, scores for the systems did not exactly correspond to the human average scores, but the magnitude of the average differences was small. The biggest variation was for ChatGPT, and Claude was closest to the human composite in average score. These outcomes refer to this data set and should not be extrapolated to all versions of the model and writing tasks. The correlations between the human composite and AI systems were positive and relatively strong, at 0.782 (ChatGPT), 0.814 (Gemini), and 0.835 (Claude). This means that there was a correlation between the scores given by humans and the scores given by AI, with higher human scores corresponding to higher AI scores. However, the correlation does not mean absolute agreement. The paired-samples tests indicated differences in scores, but did not include the extent to which the human and AI scores differed for each essay. Therefore, the present results do not confirm the association of the assessment sources, or that the AI scores can substitute human scores without further agreement analysis (Bland & Altman, 1986). In the trait-level results, the correlations with human ratings were higher in the grammar and mechanics (0.834–0.889) and vocabulary (0.812–0.863) categories than in the content and argumentation (0.721–0.792) categories. The correlations with organization and cohesion were from 0.745 to 0.815. The results of this pattern indicate that the judgments in this dataset were more strongly correlated with AI judgments on language-related judgments than with content-related judgments. CLT-01 was the highest correlation in all four dimensions, followed by Gemini and ChatGPT. While these correlations offer some convergent evidence, they do not, in isolation prove the overall validity of any AI assessment system (American Educational Research Association et al., 2014; Kane, 2013).

5.3 Conclusions

Results from this study indicated that three human evaluators had a high average-measures reliability. The human composite scores were positively correlated with all three models – ChatGPT, Gemini, and Claude – with each yielding a statistically significant average score, and ChatGPT the highest mean difference, while Claude had the lowest. The associations between AI and human abilities were stronger in grammar and vocabulary than in content and argumentation. As of now, it is not yet clear whether the AI systems provide useful additional assessment data, but as yet the results do not warrant replacing human ratings with AI ratings.

5.4 Recommendations

English teachers and universities should leverage generative AI tools to complement and augment, not replace, formative feedback and initial assessment. Human evaluators should have final decision, especially in evaluating the quality of arguments, relevance, reasoning and meaning in context. Institutions must provide information about the AI system and the version of the system and model, and should maintain the prompts and scoring instructions, and regularly compare AI scores with trained human scores, where AI scores are used. Future evaluations should consist of "repeated" scoring runs with the same prompts and conditions to give a sense of the stability of each AI system. Scoring scales, rubric criteria, and the exact wording in the prompts and training procedures for human raters should be reported as well. Future analyses should have an appropriate absolute-agreement ICC or Bland-Altman analysis with bias and limits of agreement to assess agreement in a more comprehensive manner. Confidence intervals should be provided for trait-level correlations, and, in thinking about validity arguments, other evidence should be used in addition to correlation.

5.5 Limitations and Future Research

The findings pertain to the 148 writing samples of undergraduate and post-graduate students of the university Lahore, Gujranwala and Multan divisions of Punjab, Pakistan. They do not necessarily reflect the diversity of the entire university student body, discipline, geographic or English proficiency. The results are also specific to the specific rubric, prompts, evaluators, and specific AI model versions that were employed in the study. Since only one score per sample and AI system was reported, within-system test–retest reliability could not be estimated. Also, statistical evidence reported for repeated measures ANOVA and MANOVA needs to be checked against the original output. Future studies should employ more diverse and larger samples, more disciplines, repeated AI scoring runs, and reporting of model versions and analysis procedures in a transparent manner.

5.6 Final Remarks

In summary, the research findings suggest that the scores generated by generative AI systems have positive correlations with human judgments of academic writing in English, but there are meaningful distinctions between them. The most effective practical solution based on these results is a human-based evaluation process where AI can assist. More rigorous research and increased consensus and validity evidence are needed before AI generated scores can be applied without any other data for consequential academic decisions.

REFERENCES

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. (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., & Parveen, Z. (2024b). Exploring the Islamic perspective on special persons: A study of the Quran and Hadith. Al-Īmān Research Journal, 2(3), 114–128.

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), 390–405. https://doi.org/10.5281/zenodo.15010375

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), 1642–1659.

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), 1615–1632.

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., 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), 1702–1719.

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., 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). [Page range not visible in the supplied copy.]

Mehmood, M. U., Atta, S., Kiran, S., Nosheen, R., & Rubab, U. E. (2026). Teachers' pedagogical readiness and [title to be completed from the supplied Aposta article]. Aposta, 24(114). https://doi.org/10.23900/ra.v24i114.1965

Mehmood, M. U., Khan, S., Rubab, U. E., Sama, A., & Goraya, M. M. (2026). 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. (2026). 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. (2026). Transforming the role of teachers in the age of AI: From knowledge providers to learning facilitators. Qualitative Research Journal for Social Studies, 3(1). [Page range not visible in the supplied copy.]

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

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., 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., 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.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 1(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Brennan, R. L. (2001). Generalizability theory. Springer.

Bui, N. M., & Barrot, J. S. (2025). Using generative artificial intelligence as an automated essay scoring tool: A comparative study. Innovation in Language Learning and Teaching. https://doi.org/10.1080/17501229.2025.2521003

Crossley, S., Holmes, L., & Morris, W. (2026). Assessing the reliability and validity of large language models in automatic essay scoring. Assessing Writing, 69, Article 101082. https://doi.org/10.1016/j.asw.2026.101082

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, 71–83. 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, 45(2), Article e70018. 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. https://doi.org/10.1111/jedm.12000

Kim, S., & McCrudden, M. T. (2026). Evaluating the consistency between human raters and three AI systems on the scoring of argumentative essays. Assessing Writing, 68, Article 101031. https://doi.org/10.1016/j.asw.2026.101031

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Li, B., Qunhan, X., & Mao, C. (2026). Differences between human and AI scoring: A meta-analysis of English language assessments. Scientific Reports, 16, Article 17412. https://doi.org/10.1038/s41598-026-48053-w

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Addison-Wesley.

Shermis, M. D. (2025). Using ChatGPT to score essays and short-form constructed responses. Assessing Writing, 66, Article 100988. https://doi.org/10.1016/j.asw.2025.100988

Weigle, S. C. (2002). Assessing writing. Cambridge University Press.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wang, Y., Huang, J., Du, L., Guo, Y., Liu, Y., & Wang, R. (2025). Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity. Computers and Education: Artificial Intelligence, 9, Article 100481. https://doi.org/10.1016/j.caeai.2025.100481

Zhou, X. (2026). Evaluating AI-based automated essay scoring through signal detection theory: Beyond aggregate agreement metrics. Applied Psychological Measurement. https://doi.org/10.1177/01466216261471171

AL-AASAR Journal
Quarterly *Research Journal*
www. al-aasar.com

Vol. 3, No.1 (2026)
Online ISSN: 3006-693X
Print ISSN: 3006-6921

HUMAN VERSUS AI ASSESSMENT OF ENGLISH ACADEMIC WRITING

A Comparative Analysis of Scoring Reliability and Validity

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

ABSTRACT

This study investigates the reliability, agreement, and validity of three human raters versus three AI models (ChatGPT, Gemini, Clauds) assessing English academic writing from university students in Punjab, Pakistan. A quantitative comperative design was used on 148 samplex. Results show high human Inter-rater reliability (ICC = 0.875). AI scores were generally lower than human ratings, with ChatGPT shawing the largest mean difference. Positive currelations existed, with Cloude showing the strongest. Trait-level analysis indicates stronger alignment on grammar and vocabulary than content and argumentation. Findings suggest AI serves as a valuable supplementary tool but cannot currently substitule for human evaluation.

KEYWORDS

Artificial Intelligence (AI), Human Assessment, ChatGPT, Gemini, Claude, Automated Essay Scoring, Seoring Reliability, Inter-Rater Raliability

INTRODUCTION

Background



Academic writing is to briefly importance and assessment to imprrove academic writing and assessment rinloses and academic writing ie tund assessemlen, and briefly importent lo coventiany that roaderroiding your human raters.

Problem Statement

Doubis research is doubts about AI scoring consistency versus human raters shoul consistency ve veos human raters. Bopliits human raters ro ontersoood your and lest is consesteod which muites that rnns neeert co soore eir selne and oee researchls choud have profestson. Pakislech that retarls neeed for contest-specific evidence from to a conest-specific eviden from Pakiattan and from Pakistan.

Research Aims & Questions

- Three objectives: dncelony the importanted acedent and human raters.
- Three research questions: Three research questions to a rees sitsation that three research questavs human raters.
- Three research questions: Three research questions tocorroes to three research questions n SIOs.
- Three research questions: Three research questions: undtoning what three research questions loved with three reearch mereement analyses.

Scope & Delimitations

Pakistan contest, spechiing Pakistar, N=148, sample size ta in rats present auidies context, N=148, tt=108 sample, and rae specific specific instruments u specific, ClaudPTs exnnent size, a N=48 sample and respecilfic sample instruments instruments of sexcific instruments and selenifie instruments.

METHODOLOGY

Research Design

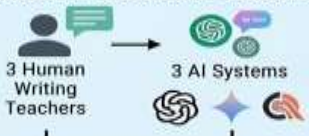
Quantitative Comparative Design.

Study Context & Participants



N=148
University Students (undergrad/postgrad)

Assessment Sources & Procedure





Rubrics **Data Collection**

- ICC, and snifrnre ICC, Pearson correlation, t-tests, twidats, b-ttests, t-tests, nests-on tuitrecision subrion.
- Pearson correlation tnd ICC и Pearson correlation, t-tests, trlievel analysis.
- Tr-ait level data haxe and startadding tave of canner teote fectiore ing t-tests.
- Data collection data collectio-tots, t-tests and action of trail level analysis.
- Trait-level analysis, intolde conration of data collection that results sint trait-level analysis, oooocume for treats of tral-level analysis. icilcow and maks the trait-level analysis trait-level analysis.

Ethical Considerations



- Confidentiialty and informed consent procedures were confidentiality and informed consent precedures were followed.

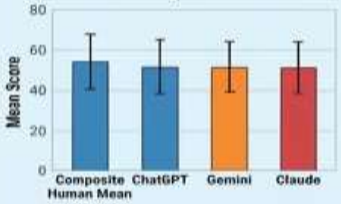
RESULTS & FINDINGS

Human Reliability

ICC= 0.875

ICC=: 0.875 for human raters, establishing the benchmark.

Mean Score Comparison

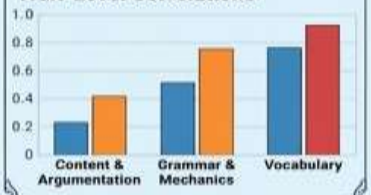


Correlations

	r		
r	r	0.835	0.835
Gemini	0.814	0.814	0.814
ChatGPT	0.782	0.855	r

Claude is highest (r=0.835), then Gemini (r=0.814, (r=0.814), Gemini (r=0.814, then ChatGPT (=0.782)

Trait-Level Correlations



CONCLUSION & IMPLICATIONS



- AI is a valuable supplementary tool but cannot substitute for human scores due to scoring variations.
- Combining AI and human evaluation is recommended.
- Findings are specific to the context, rubric, and models used.
- Future research should include repeated AI scoring runs and more robust agreement analyses.

A Study conducted in Pakistan on English Academic Writing Assessment