AI or Peer Feedback: What Works Best in Improving Writing?

Saira Mahmood
Department of Education

University of Oxford

Dissertation submitted in part-fulfillment of the requirements for the degree of Master of Science in Applied Linguistics for Language Teaching.

Abstract

Feedback is a key driver of student learning, yet teachers often struggle to provide detailed written corrective feedback (WCF) at scale. Alternatives such as peer review and automated writing evaluation (AWE) have therefore gained attention, but their relative effectiveness remains underexplored, particularly in low-resource contexts. This study investigates whether AI- or peer-generated feedback leads to greater improvements in writing, whether the two differ in scoring and the types of feedback they produce, and how students perceive and evaluate them.

The research involved 36 female undergraduates in a B.Ed. program in Karachi, Pakistan, randomly assigned to AI feedback (n=12), peer feedback (n=11), or control (n=13) conditions. All participants wrote an initial essay, reviewed a peer's draft, and revised their own essay either in light of feedback (AI or peer) or no feedback (control). Both drafts were scored on a 54-point rubric across three domains: content and organisation, language use, and mechanics. Gain in scores across drafts were calculated. Quantitative analysis (ANCOVA) was complemented by a survey and independent coding of feedback quality. Participants language proficiency was accounted for.

Results showed that AI feedback yielded the largest mean gains (M=5.0), followed by peer feedback (M=3.9) and control (M=2.9), though differences were not statistically significant. Importantly, AI feedback produced a significant interaction with proficiency in mechanics: less proficient students showed greater gains when revising with AI. Qualitative data revealed that AI comments were more directive and elaborated, while peer feedback leaned toward praise and surface-level suggestions. Students found AI comments more comprehensive and accurate.

Acknowledgements

I am thankful for the support and help I have received throughout my degree in Applied Linguistics. My supervisor, course instructors, peers, and colleagues have been invaluable sources of strength on what was a long and deeply rewarding journey. This course has made me a better educator by broadening my horizons, and introducing me to previously unknown fields of study.

My heartfelt gratitude goes to Dr. Elizabeth Wonnacott, my supervisor, whose insight, expertise and critical engagement with this work shaped it in crucial ways. I am thankful for the grace and patience she has showed me over the year as I balanced health concerns and work commitments alongside this dissertation. Thank you also to Dr. Hamish Chalmers for his commitment to this course. A special note of thanks is reserved for Adele Gregory, the Course Administrator for the Applied Linguistics programme, who worked tirelessly to ensure students had the support and guidance they needed to be successful.

I am grateful to my cohort members whose assistance and companionship made this degree feel anything but 'distance learning.' Thank you to Dr. Hameedah Sayani for all her work during the data collection process of this study.

I owe to my parents a tremendous debt of gratitude for their support of me in all my endeavours. Their patience and steady reassurance gave me the strength to persist through moments of doubt. I dedicate this achievement to them and to Ms. Salma Ahmed Alam, whose vision and commitment to transforming the educational landscape in Pakistan have been deeply inspiring. As my mentor, she not only believed in my potential but also worked tirelessly to secure the funding that made this degree possible.

Finally, thank you to Qasim, Ibrahim, and Fatima, for their understanding during the many hours that their aunt had to be absent, and for excusing (albeit begrudgingly) my absence from our weekly game nights.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

WCF	Written Correct Feedback
L1	First Language
L2	Second Language
AI	Artificial Intelligence
SRL	Self-Regulated Learning
AWE	Automated Writing Evaluation
EFL	English as a Foreign Language
ESL	English as a Second Language
GenAI	Generative Artifical Intelligence
AP	Advanced Placement
EAP	English for Academic Purposes
CUREC	Central University Research Ethics Committee
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
MANCOVA	Multivariate Analysis of Covariance

Chapter 1: Introduction

This study compares the effectiveness of peer and AI feedback in improving students' writing outcomes in a Pakistani English as a Foreign Language (EFL) higher education context. Feedback can broadly be defined as any evaluative or corrective information on their task performance that learners generate or obtain from multiple sources (see Hattie & Timperley, 2007; Smith & Lipnevich, 2018). In particular, this study focuses on Written Corrective Feedback (WCF), in the form of scores and/or comments assigned to student writing. Importantly, WCF can refer both to feedback provided on the global level (e.g., content, organisation, style, etc) or local level (i.e., grammar, punctuation).

1.1 Rationale

In education, feedback is regarded as a highly impactful and crucial component of the teaching and learning cycle, with a largely positive effect on student learning (Black & William, 1998; Hattie & Timperley, 2007; Shute, 2008; Hyland & Hyland, 2006; Wisniewski et al., 2020; Hattie, 2009). The Education Endowment Foundation's review of extensive research concluded that feedback could have a "very high impact for very low cost based on extensive evidence" (EEF, 2021). Nevertheless, feedback is easy to get wrong. Roughly a third of research studies on feedback have found that it can negatively influence learning (Bangert-Drowns & Kulik, 1991; Kluger & DeNisi, 1998). Critical feedback can to lead to feelings of shame and anger for students (Ryan & Henderson, 2017), and excessive praise may cause task performance to stagnate (Ende, 1983). Therefore, any feedback interventions must be planned carefully.

In the classroom, the most common feedback agents are teachers. While usually effective (Lv et al., 2021), teacher feedback is time-consuming, and its efficiacy may decrease in low-resource large-class-size settings (Kadek et al., 2022; Allen et al., 2021). Additionally, having to balance giving large amounts actionable feedback with managing student emotions can be emotionally taxing for teachers (Yu et al, 2021). As a result, alternative sources of feedback must be considered which may supplement teacher feedback in classroom and so lighten teacher's workload. This study aims to compare the effectiveness of two such sources: peers and AI.

1.1.1 Peer Feedback

Peer review positions students as both writers and evaluators, creating opportunities for audience awareness, self-regulation, and transfer of revision strategies. Its strengths are practical and pedagogical: it distributes feedback work across the class, provides multiple perspectives, and can foster motivation through dialogue. However, peer feedback is highly variable in quality. Without scaffolds, comments may skew toward praise

or vague advice; in L2 contexts, concerns about linguistic accuracy can undermine trust (Adams et al., 2011; Philp et al., 2010; Tsivitanidou et al, 2011; Van de Weghe, 2004)

Research consistently shows that reliability and usefulness improve when peers work with clear rubrics and brief training (Van Steendam et al., 2010; Cho et al, 2006; Panadero et al, 2013). This study makes use of a detailed rubric and feedback template to compare the effectiveness of peer review with AI review in a Pakistani EFL setting with participants who are student-teachers and so have been trained in giving feedback.

1.1.2 AI Feedback

Generative AI chatbots can deliver immediate, criterion-aligned feedback at scale, with particular strengths in coverage, consistency, and attention to local language issues (McCurry, 2010; Li et al., 2015). When guided by detailed prompts and rubrics, AI can produce specific suggestions and metalinguistic explanations that are time-intensive for teachers to provide routinely. Yet AI feedback has limits: it can be overly generic or formulaic, miss deeper argumentative issues, or overwhelm learners with volume (Stevenson & Phakiti, 2019; Lin & Crosthwaite, 2024). Its effectiveness also depends on prompt design and user competence, raising questions about equity and classroom implementation (Fleckenstein et al., 2023). In institutional settings, teacher mediation, transparent use policies, and alignment with assessment criteria are essential to convert AI's speed and breadth into learning value. This study therefore evaluates AI feedback under controlled, rubric-based conditions that are feasible for instructors to implement.

1.2 Research Questions

The rationale has led to the development of the following research questions

RQ 1 Does the type of feedback received (AI, peer, or none) significantly affect participants' overall improvement, as well as their domain-specific gains in content, language use, and mechanics across two drafts, while controlling for initial language proficiency?

RQ 2. Do AI and peer reviewers assign significantly different overall or domain-specific (Content, Language Use, Mechanics) scores to participants' first drafts when using the same rubric?

RQ 3. How do AI- and peer-generated feedback differ in participants' perceptions and attributions and independently coded quality characteristics?

1.3 Overview

This dissertation has been divided into five sections: introduction, literature review, methodology, results, discussion and conclusion.

Chapter 2 reviews scholarship on written corrective feedback (teacher, peer, and AI), highlighting conditions under which feedback is most effective and gaps in direct AI–peer comparisons in EFL contexts.

Chapter 3 details the context, participants, instruments, procedures, and analytic approach for the mixed-methods design.

Chapter 4 reports quantitative results for improvement and scoring alignment, followed by qualitative analyses of students' perceptions and coded feedback characteristics.

Chapter 5 discusses the findings thematically, outlines implications for practice (including how AI and peer review can supplement teacher feedback to ease workload), and notes limitations with recommendations for future research. Appendices provide instruments, prompts, and additional tables/figures.

Chapter 2: Literature Review

2.1 Feedback: Overview and Definition

Feedback is a crucial component of education and has been linked to both improved student learning outcomes and enhanced student satisfaction and motivation (Hattie & Timperley, 2007; Hyland & Hyland, 2006; Wisniewski et al., 2020; Espasa & Meneses, 2009; Narciss & Huth, 2004; Lou & Noels, 2020). However, feedback can also negatively effect students' learning or affective state, especially if it is too long, not tailored to students' proficiency levels, too critical or uses excessive praise (Ryan & Henderson, 2017; Ende, 1983; van Merriënboer and Sweller, 2005; Lu et al., 2023; Daumiller & Meyer, 2025)

To make sense of the varied effects of feedback and to identify what type of feedback is most effective, it is first necessary to consider how feedback itself has been defined. In education, three broad purpose-driven definitions of feedback have been considered: feedback as motivation, reinforcement, or information (Nelson and Schunn, 2009, p.379).

The first two definitions are rooted in behaviorist assumptions, which hold that behaviours followed by satisfying consequences are more likely to be repeated, while those followed by unsatisfying consequences are less likely to recur (see Skinner, 1968; Thorndike, 1927; Kulhavy, 1977). However feedback cannot be reduced to a simple mechanism of reward and punishment, as both praise and criticism can have unintended consequences on learners' emotions and subsequent task performance (Ende, 1983; Ryan & Henderson, 2017). Recognizing the central role learners themselves play in the effectiveness of feedback, constructivist theories of teaching and learning reconceptualized feedback so that it was no longer information transmitted by an expert (e.g., a teacher), but was understood as evaluative or corrective information on learners' task performance that learners could generate or obtain from multiple sources—teachers, peers, books, parents, personal experience, self-monitoring, or even natural consequences (Eraut, 2007; Hattie & Timperley, 2007; Smith & Lipnevich, 2018; Sadler, 1989; Butler & Winne, 1995).

While feedback can take on many forms in the classroom, this review will focus mainly on written corrective feedback (WCF), in the form of scores and/or comments assigned to students on a written production, such as an essay. It is important to remember that WCF can refer both to feedback provided on the global level (e.g., content, organisation, style, etc) or local level (i.e., grammar, punctuation) (Al-Jarrah, 2016; Crosthwaite et al., 2022).

While teachers are the primary agents of WCF, teacher feedback is time-consuming and less effective in large classes (Allen et al., 2021). As a result, a growing body of research has examined the role of alternative feedback sources—such as peer feedback or automated feedback evaluation (AWE) systems.

The research on the effectiveness of such alternative feedback sources has yielded mixed results. In some cases, teacher feedback has been found to be more effective (Lv et al., 2021) and more preferred by students (Maas, 2017; Yang et al., 2006) than other forms of feedback. In fact, students may sometimes doubt that peers possess the linguistic competence to give them good feedback (Adams et al, 2011, Philp et al, 2010). However, in certain contexts, peer or AWE feedback agents have been shown to complement or even substitute for teacher feedback (e.g., Double et al., 2020; Latifi & Noroozi, 2021; Gielen et al., 2010; Wang & Han, 2022; Zhang, 2020; Huawei & Aryadoust, 2023).

To clarify the practical benefits of peer and AWE feedback on student writing, this section is organized in two parts. It first reviews the major taxonomies of WCF comments that have been proposed in the literature, outlining how different types of feedback influence revision. It then evaluates empirical research on peer and AWE feedback systems, considering their relative effectiveness and the contexts in which they complement or substitute for teacher feedback. Since a limited number of studies compare AI and peer feedback directly, the sections on the respective effectiveness of each are important.

2.2 Written Corrective Feedback: Typologies and Effectiveness

2.2.1 Typologies

Written corrective feedback (WCF) refers to scores and/or comments assigned to students on a piece of their writing (e.g., an essay). WCF can address both the local aspects (e.g., writing mechanics) and global aspects (e.g., content, organisation, or language use) of writing (Al-Jarrah, 2016; Crosthwaite et al., 2022).

Over the years, multiple taxonomies of writing features have been proposed by scholars and practitioners (Knoblock & Drake, 2005; Gentry et al., 2014). These have been instrumental in tailoring feedback comments and scores to particular aspects of writing and have been used to create evaluative rubrics. For example, the popular Six Traits of Writing (Kozlow & Bellamy, 2004) rubric is based on a framework proposed by Paul Diederich in the 1960s and independently verified by Spandel (2012) and Education Northwest (Coe et al., 2011). These Six Traits are ideas, organisation, voice, word choice, sentence fluency, and conventions. Proceeding from taxonomies of written material, writing feedback has often been characterized as an aggregate of its global (holistic), discourse (content and organization), and form (grammar) concerns. Table 1 reproduces these subcategories from Pearson (2022).

Table 1: Levels, Categories, and Sub-categories of Textual Features for Written Feedback ; Commentary Response

Level	Category	Sub-category
General	Overall quality of essay in all its aspects	
Discourse	Content	Clarity or understandability
		Development or lack of development
		Overall quality of content
		Accuracy of information, truth value of a claim, accuracy of interpretation
	Organization, coherence, cohesion	Transitions
		Thesis statement
		Topic sentence
		Overall quality of organization
		Coherence, cohesion
		Idea placement
		Paragraph order
Form	Vocabulary	Word choice, collocation, phrasing
		Overall quality of vocabulary
	Grammar/Syntax and morphology	Sentence structure
		Omission
		Word order
		Verb tense or form
		Noun form
		Article
		Agreement
		Preposition
		Pronoun
		Overall quality of grammar
Form	Mechanics	Punctuation

	Spelling
	Documentation or attribution
	Formatting and style
	Overall quality of mechanics

Note. From "A typology of the characteristics of teachers' written feedback comments on second language writing" (Table 2), by W. S. Pearson, 2022, *Cogent Education, 9*(1). © 2022 The Author(s), CC BY 4.0.

Although taxonomies of writing comments have been instrumental in shaping feedback frameworks, typologies focused specifically on WCF as a whole—for example, those addressing feedback in the form of underlining, marginal comments, or error codes, alongside comments—are relatively scarce. Moreover, existing research in this area contains inconsistencies that complicate efforts to draw firm conclusions about the effectiveness of different forms of WCF (Evans et al., 2011; Bonilla López, 2021). This is primarily because WCF has been characterised heterogeneously (Brown et al., 2023). According to Comajoan-Colomé & Salguero (2024), WCF typologies have emerged from two distinct form of research: conceptual and empirical studies. Conceptual typologies derive from broader theoretical attempts to categorise WCF, whereas empirical typologies typically emerge as by-products of experimental studies in which researchers code the feedback found on student writing and classify it according to emergent categories.

Appendix G offers a selective summary of conceptual WCF typologies.

Appendix H offers a selective summary of empirically driven WCF typologies.

2.2.2 Effectiveness by Type

These typologies underscore the heterogeneity of WCF practices and serve as a framework for analysing the effectiveness of different types of corrective comments or marks. Notably, early research was sceptical of WCF, with some scholars arguing that it was not only ineffective but could even be harmful to learners' language development (e.g., Krashen, 1982, 1985; Truscott, 1996, 1999, 2007). However, most early research studies restricted the use of the term WCF to grammar errors.

Nevertheless, some studies have indicated that certain types of WCF do not lead to substantial gains, at least under certain conditions. Fazio's (2001) classroom experiment with 112 Grade-5 francophone and French-as-second-language writers found no accuracy improvements under different WCF conditions (corrections, content comments, or their combination) and most pupils consulted feedback only sporadically. While this study has been criticised for its small sample size and lack of control group, other studies have also found similar results. For instance, Trustcott and Hsu (2008) found that while indirect WCF (where errors are

flagged but no correction offered) did result in accuracy gains for the feedback-receiving group for the same text, there was no significant difference in the error rate for the control and experimental groups over the long-term (i.e, on a new piece of writing).

In contrast to these accounts, however, a growing body of empirically driven WCF research (see Appendix H) has identified specific types of feedback comments that may be effective in promoting accuracy. For example, multiple studies with both native students or English as a Second Language (ESL) learners have found that while both direct (error flagged, correct form offered) and indirect (error flagged, no correct offered) corrective feedback types are effective for short-term writing proficiency gains, direct WCF is more effective for long-term gains (Van Beuningen et al., 2008; Van Beuningen et al, 2012). Furthermore, direct feedback with meta-linguistic explanations, where the reviewer gives a clue about the nature of the mistake either via error codes or brief descriptions, is most effective for long term grammar gains (Bitchener & Knoch, 2010). Studies comparing feedback where the teacher corrects all types of linguistic errors (unfocused) with feedback where the teacher focuses on correcting a particular error (focused) found that both approaches lead to greater improvement than no feedback for learners in EFL contexts (i.e., those studying English as a foreign language in settings where English is not widely spoken; Ellis et al., 2008). By contrast, in ESL contexts (i.e., learners studying English as a second language in English-speaking environments), focused feedback produces stronger gains than unfocused feedback (Sheen et al., 2009).

Despite early scepticism, over two decades of empirical studies and meta-analyses show that WCF is generally effective under the right conditions. Graham et al.'s (2015) meta-analysis found small-to-large effect sizes ($d = 0.38$ to $d = 0.87$) for writing feedback, while Biber et al. (2011) showed that feedback addressing both content and form produced greater gains in grammatical accuracy than feedback limited to form. Beyond gains in writing proficiency, WCF has been found to sharpen writers' self-evaluation skills (MacArthur, 2007), facilitate interlanguage development (Bonilla Lopez et al., 2018), and support broader L2 acquisitional processes such as text editing and noticing (Truscott & Hsu, 2008; Storch & Wigglesworth, 2010).

When WCF is direct, targeted, focused on particular errors, includes a metalinguistic component, and is directed at sufficiently proficient learners, it can lead to substantive improvements in writing, at least in the short term (Sheen et al., 2009; Kang & Han, 2015).

However, much of this evidence for WCF effectiveness comes from studies measuring improvement within the same draft on which teachers provided feedback, rather than in future pieces of writing, making it difficult to draw firm conclusions about long-term gains. Where studies have measured gains on a new piece

of writing, it has been found that direct WCF (where corrections are explicitly provided) is more strongly associated with long-term gains when compared to indirect feedback (where errors are simply identified) (Bitchener & Knoch, 2010; Truscott & Hsu, 2008; Van Beuningen et al., 2008; Van Beuningen et al, 2012).

2.2.3 Effectiveness by Proficiency

Learners' L2 (second language) proficiency is a salient but relatively underresearched factor shaping the reception of WCF (Allen & Katayama, 2016). Kang and Han's (2015) meta-analysis of 22 studies demonstrated that WCF has an overall moderate-to-large positive effect on L2 writing accuracy (Hedges's $g = .68$, $p < .0001$), though the efficacy varied across conditions ($Q = 87.18$, $p < .0001$). Their analysis of nine moderators indicated that proficiency plays a central role: beginners benefited little ($g = .09$), whereas intermediate ($g = .56$) and advanced learners ($g = .74$) showed substantial gains. Scope and type of feedback also mattered: focused feedback was more effective than unfocused ($g = .598$ vs. $g = .361$), and direct feedback was more effective than indirect ($g = .69$ vs. $g = .329$). Against this backdrop, Paris (2020) investigated whether proficiency moderated the effectiveness of direct versus indirect WCF among 7 undergraduate learners of German at a Canadian university (four beginners, three advanced). She found that proficiency level did not influence accuracy gains: both groups' error rates increased after direct feedback but decreased after indirect feedback. However, her small sample size limits generalizability. Crucially, neither Kang and Han (2015) nor Paris (2020) investigated whether the effects of feedback differed by proficiency level within specific writing subcategories (e.g., content, mechanics; see Table 1).

2.3 Feedback Type and Effectiveness Across Agents

While much of the research on the effectiveness of WCF has focused on teacher feedback, providing detailed written feedback is often time-consuming for teachers and difficult to sustain in large classes. These practical constraints have led to growing interest in alternative feedback agents, particularly peers and automated writing evaluation (AWE) systems, with a number of studies focusing on these (e.g., Double et al., 2020; Latifi & Noroozi, 2021; Gielen et al., 2010; Wang & Han, 2022; Zhang, 2020; Huawei & Aryadoust , 2023).

2.3.1 Peer Feedback

According to Wu & Schunn (2021), peer feedback is more likely to prompt revisions when it explicitly identifies errors and provides actionable solutions. Other studies lend credence to this idea. Nelson and Schunn (2009) investigated how peer feedback characteristics shape undergraduate students' writing revisions, focusing on how features of the feedback interacted with writers' internal mediators (e.g., understanding or agreement) and their likelihood of implementing changes. Each participant essay was reviewed by six peers following structured guidelines (e.g., clarity of prose, strength of argumentation). In

total, over 1,000 feedback segments were analysed and first grouped as problem/solution, praise, or summary. Within the problem/solution category, feedback was further coded by whether it identified a problem only, gave a solution only, or included both; whether it addressed global (e.g., "All of your arguments need more support") or local (e.g., "You used the incorrect form of 'there' on page 3. You need to use 'their'") issues; whether it included affective language such as mitigation-compliments (e.g., "Your main points are very clear, but you should add examples"). Implementation was measured by comparing first and revised drafts to see if feedback was incorporated. The results showed that explicit solutions were the only feature that significantly influenced revision (p=.03). Among the mediators, understanding the problem was the key predictor of implementation (p=.04), whereas simple agreement was not. Features that enhanced understanding included clear localization (p=.001), summaries (p=.05), and explicit solutions (p=.06). It was also noted that extensive explanations of the problem reduced understanding (p = .03).

Problematically, direct, actionable suggestions are comparatively rare in peer feedback. For example, Cho et al. (2006) found that expert feedback was more than three times likely to contain directive comments (comments suggesting specific changes) when compared to peer feedback. Peer feedback contained more than 70% praise when compared to expert feedback. In analysing comments by students and himself (a teacher), Caulk (1994) found that his comments tended to identify many errors whilst peers focused only on some errors, and that his comments were 8% more focused on form when compared to peers'. Studies in Asian EFL contexts have found that peers tend to provide global/holistic comments, while teachers or writing experts focus more on local/form-based issues (Chen, 2010; Xu & Liu, 2010).

Some ways to increase the amount of direct and mechanics-based feedback in peer review comments are training, the use of rubrics which outline a solid evaluative criteria, and having multiple peers give feedback to a single essay (Cho et al, 2006; Panadero et al, 2013). To investigate the benefits of multiple peer reviews, Cho and McArthur (2010) randomly assigned 28 undergraduate students to three feedback conditions: feedback from one expert (SE), one peer (SP), or many peers (MP). The researchers' goal was to investigate what type of written revisions emerged within each feedback condition group. Students drafted a short report and used a rubric to review six classmates' draft. Meanwhile, an expert also reviewed every draft under double-blind conditions. Students then revised their essays based on the feedback received (SE vs SP vs MP). The feedback messages for each condition were coded according to Cho et al (2006)'s typology: directive (a specific change suggested), non-directive (a problem flagged, or broad direction given without a specific solution), praise (positive evaluation), criticism (negative evaluation sans solution), summary (recap of main points), and off-task (unrelated remarks). An analysis revealed that MP condition participants received more comments of every type. SE feedback comments were 83% direction. Participants in the SE feedback

condition made the most simple repairs (mechanical surface-level fixes) ($p = 0.005$), while their MP group counterparts writers made more complex repairs (micro-level meaning changes) ($p = .012$) and extended content revisions (elaborations and/or justification of existent points) ($p = .026$). Directive comments were positively associated with simple repairs, whereas non-directive (comments predicted complex repairs ($\beta = 0.43$, $p = .021$) and new content additions (e.g., new points, including entire new paragraphs) ($\beta = 0.42$, $p = .031$). Praise had no effect.

Peer reviews also benefit from the use of rubrics and feedback training. In the absence of a well-defined evaluative criteria, peer reviewers may resort to comments that offer empty praise (Tsivitanidou et al, 2011; Van de Weghe, 2004) or provide surface-level corrections (Simmons, 2003). Explicitly training peers in how to give good feedback has been shown to increase the higher-order comments provided by peers (Van Steendam et al., 2010). Cui et al (2021) found that, after training, peers could provide meaning-focused comments that are comparable to comments produced by teachers in high workloads enviroments (2022). Falchikov & Goldfinch (2000)'s meta-analysis of forty-eight quantitative peer assessment studies comparing peer and teacher marks found that peer marks resemble teacher marks more closely when peers grade using well understood criteria. Schunn et al. (2016) examined whether AP English students could reliably and validly assess peers' essays when guided by a student-friendly rubric. In the study, 1,215 students across 26 schools wrote AP-style essays, then reviewed five peers' essays anonymously using the rubric. Their ratings were compared with teacher scores (489 essays rated with the same rubric) and expert AP scorers' scores (100 essays rated with the official AP holistic rubric). Reliability was measured with interclass correlations (ICCs) across student ratings (all > .40 except conventions), and validity by correlations with teacher and expert scores ($r \approx .40–.70$ with teachers; $\geq .50$ with experts). The average of five peer ratings was slightly more valid than a single teacher rating, showing that carefully designed rubrics make peer feedback both reliable and comparable to expert judgments.

While these studies underscore the positive effects of rubrics and training on peer-assigned scores, they do not directly offer insights into whether rubrics and training also help peers give more direct comments or produce feedback messages matching those of teachers'. In a study involving undergraduate history students, Patchan et al. (2009) compared end comments (general feedback comments given at the end of an essay) by students with end comments by a writing instructor and a course instructor on student essays. Over 1400 comment segments were analysed. The researchers found the comments to be fairly similar across the three reviewers. While the study did not explicitly set out to measure the effect of a rubric and so lacks a control (no rubric condition), the researchers did note that the comparable quality of peer comments was likely due to the presence of a rubric and incentives to take peer review seriously (p. 143). Given that peer

review is crucial to decreasing teacher workload and offering better educational opportunities to students, more research is needed to investigate how the use of rubrics and training impact the type of feedback comments produced during peer review.

2.3.2 AWE Feedback

Another source of WCF are Automated Writing Evaluation (AWE) systems, which analyse and score written work via computer programmes based on a combination of computational linguistics and artificial intelligence (Shermis & Burstein, 2003). In addition, recent GenAI chatbots are powered by large language models that generate adaptive, human-like responses through deep learning. Increasingly ubiquitous, AI-powered chatbots can enhance students' writing proficiency and capacity for self-regulated learning (SRL) capacity by offering them immediate and personalised WCF (Ranalli & Yamashita, 2022; Molenaar, 2022; Kohnke et al., 2023).

Studies investigating the effectiveness of AWE systems in improving writing proficiency show mixed results. AWE systems have been praised for increasing learner autonomy and engagement (Wang et al, 2013; Weigle, 2013), providing immediate and personalised feedback (Dikli, 2006; Ranalli & Yamashita, 2022, Henderson et al., 2015), and decreasing teacher workload (Langove & Khan, 2024). Other studies have found that AWE systems are better at identifying local low-level errors than they are offering feedback on more global concerns such as content and organisation (McCurry, 2010; Stevenson & Phakiti, 2019, Wang, 2020; Lang et al., 2019). Li et al (2015) notes that while students using the AWE-system Criterion® reported high-levels of general satisfaction, they also felt that the feedback model lacked specificity in its comments on content and organisation. However, Ngo et al. (2024)'s three-level meta-analysis which examined 58 between-group and within-group studies investigating the effectiveness of AWE in EFL/ESL writing reported a different finding. Between-group studies (n = 24) compared the writing outcomes of students receiving AWE support to the writing outcomes of students receiving teacher and/or peer feedback. An analysis of between-group studies revealed that AWE feedback had a small effect (g = 0.27) on grammar and mechanics. The authors attributed this to several studies relying on less efficient AWE systems such as Pigai and Jukuu.

Previous studies have indicated that direct and metalinguistic feedback that has been associated with gains in learner's writing proficiency (Sheen et al., 2009; Kang & Han, 2015; Bitchener & Knoch, 2010;Van Beuningen et al., 2008; Van Beuningen et al, 2012). Research specifically focused on feedback by generative AI chatbots suggests that these are the type of comments chatbots most often produce, atleast for issues of grammar and mechanics. In contrasting teacher feedback with ChatGPT (GPT-4) feedback, Lin & Crosthwaite (2024) found that GPT's direct fixes were more local (i.e. grammatically oriented) than global

(i.e. content-oriented) when compared to teachers (0.93 vs 0.84). However, they were also slightly more accurate (0.987 vs 0.949). GPT exceeded teachers in reformulation (offering sentence rewrites) and metalinguistic (information about errors or rules) feedback. Nevertheless, GPT also seemed to offer more superfluous or unnecessary advice for these errors, which is something that future studies can guard against by offering GPT a feedback template or asking it to restrict its responses to a specific word count. Notably, this experiment did not include a rubric, but the prompt given to ChatGPT and the teacher noted that the text provided to them "has a number of issues with grammar, vocabulary, organization, and ideas" and they could "provide written corrective feedback on this work in any form [they] choose".

For AWE systems to produce feedback comparable to teachers, effective prompting seems to be necessary. Yu and Xie (2025) conducted a controlled study comparing ChatGPT- and teacher-generated WCF in an EFL high school context in China. 60 lower-intermediate-English level students wrote two argumentative essays in a counter-balanced design, alternately receiving feedback from ChatGPT or their teacher, then revising and resubmitting. Four teachers (one class teacher, and three other teachers) and ChatGPT provided feedback on 240 essays, generating 1,200 records for analysis. The researchers ensured that ChatGPT and the four teachers were all given a similar prompt which asked them to "give feedback" on students' essays at "the surface and meaning levels" before defining each of these and explicitly asking for correct surface-levels forms and meaning-level revision suggestions for each error found in the review. The surface-level feedback received was coded as Spelling, Singular-Plural, Verb Tense, Subject-Verb Agreement, Articles, Pronouns, Collocation, and Punctuation and the meaning-level feedback was coded as Ideas and Elaboration, Organization, and Logic. The feedback points were counted. Findings showed that ChatGPT and the class teacher had the highest correlation among all possible reviewer pairs within the five reviewers ($r=.483$, $p<.001$). In contrast, highest correlation among the human teachers was $r=.404$ (between Teacher 2 and Teacher 3). ChatGPT's surface-level feedback (grammar, spelling, mechanics) was as comprehensive and accurate as that of the class teacher, though less extensive than feedback from the other three teachers. Importantly, ChatGPT produced more meaning-level feedback (content, organization, logic) than some teachers, often with detailed explanations and examples. The researchers note that ChatGPT's surface-level feedback offered metalinguistic explanations of correct errors (e.g., "'impact' is a countable noun, so you could use 'impacts' instead of 'impact' in the first sentence"). Meanwhile, teachers were more likely to offer indirect feedback (the identification of errors with correction) or supply corrections sans further explanation. Teacher feedback, however, while shorter, was pitched to students' proficiency, and easier to act upon; ChatGPT occasionally used advanced or unfamiliar vocabulary (e.g., it recommended the use of the word "ubiquitous" which was new to the students). Students' uptake of teacher feedback was slightly

higher (78.1%) than that of ChatGPT (70.4%). Uptake of surface-level feedback was high for both sources (≈87–92%), while uptake of meaning-level feedback remained low (≈29–32%). Revision strategies were similar across conditions: correction dominated surface-level revisions (Teacher: 66.6%; ChatGPT: 61.5%), while "no correction" was the most common response to meaning-level feedback (>60%). The researchers do not explicitly note the version of ChatGPT they used, although they mention that GPT-4o was the latest version released around the time of their experiment. In addition to corroborating earlier findings that GPT offers direct and metalinguistic feedback, this study also underscores that effective prompting can allow GPT to produce meaning-level feedback comments.

In addition to the differences in feedback comments from teachers vs those from AWE systems, there have also been differences noted in the scores assigned to student writing by the two agents. Li et al. (2015) found low to moderate correlation between AWE scores and instructor scores in their university-based study involving three ESL writing instructors and 67 students, suggesting that AWE systems are inefficient scorers. However, Keith (2003)'s review of AWE validity studies noted that correlations between AWE systems and human raters were very similar to those found between human raters themselves. This discrepancy may be because this review, like early summary papers on the topic, mostly analysed score correlations across large-scale standardized. Interrater reliability can be expected to be lower in classroom contexts where the content of student writing is likely to be more important (Warschauer and Grimes, p. 24). Notably, while García-Varela et al. (2025) have found that ChatGPT's grading becomes more stable (scores remain steady across identical responses) and fair (criteria applied evenly to all students' work) when guided by a detailed rubric, more research into the reliability of AI-scoring is needed.

In addition to scoring concerns, it has been found that AWE feedback does not benefit all students uniformly. Fleckenstein et al. (2023)'s multi-level meta-analysis on the effects of AWE feedback on student writing in English L1 and L2 classes found that AWE feedback leads to moderate improvement, but large and unexplained variances suggest that learning characteristics still play a huge role. Yan and Zhang's (2024) study with AI chatbots as feedback tools found that learners thought that ChatGPT feedback was highly accurate, more engaging and less intimidating than teacher or peer feedback. However, the benefits of AI feedback were mediated by students' language proficiency and digital competence. Higher-proficiency, tech-savvy students were able to craft effective prompts and strategically interpret revisions, while their lower-proficiency peers benefited less and tended to use AI feedback less effectively. Also, students across proficiency levels reported that generating prompts and interpreting feedback was cognitively demanding and mentally taxing. Perhaps due to these limitations, practitioners mostly use AWE systems alongside human raters, as this combination can maximize the strengths of both (Ferreira et al., 2007; Potter & Fuller,

2008; Warschauer & Grimes, 2008). In this hybrid setup, the negative effects of AI use could perhaps be lessened if teachers served as intermediaires between students and AI feedback systems, as students would no longer have to directly interact with taxing AI systems,

2.3.3 Peer and AI Feedback Comparison

There are relatively few empirical studies comparing AI-driven AWE feedback with peer feedback and more research is needed (Shi & Aryadoust, 2024). Some of the available comparisons indicate mixed results. Bergström & Yvdal (2024) found that ChatGPT produces more detailed and constructive feedback when compared to peers. Meanwhile, Fang et al. (2025) has noted no significant differences in the effectiveness of peer and AI feedback.

In a study involving 30 EFL students in Israel, Zeevy-Solovey (2024) found that peer feedback primarily targeted organization and mechanics, ChatGPT addressed a broad range of components (especially grammar, content, and mechanics), and teacher feedback focused heavily on grammar and mechanics but also covered expression and content. However, since feedback in this study was given to students sequentially (peer → ChatGPT → teacher), earlier revisions may have influenced the scope of later feedback, with some issues potentially already resolved before teacher or AI review.

Banihashem et al. (2024) investigated whether feedback on argumentative essays differed when provided by peers versus ChatGPT, and whether the quality of the original essay predicted the quality of feedback from either source. Seventy-four Dutch university students in the life sciences wrote an argumentative essay on one of three controversial food-science topics and then gave two digital peer reviews. Using the same prompts as the students, ChatGPT also generated feedback on each essay. Essay quality was rated by two researchers, who also categorized feedback comments based on Kerman et al. (2022)'s feedback taxonomy. Feedback segments were coded as Affective (positive or negative emotion such as praise), Cognitive–*Description* (providing a summary of the essay); Cognitive–*Identification* (pinpointing a specific problem); Cognitive–*Justification* (explaining why a problem may have occurred); Constructive (offering an actionable recommendation). After coding each comment, the researchers scored them from 0 (poor quality) to 2 (good quality). Results showed that ChatGPT produced significantly richer descriptive summaries of essays ($M = 2.00$ vs. 1.91, $p < .05$), whereas peers were better at pinpointing problems ($M = 1.52$ vs. 1.29, $p < .01$); affective, justification, and constructive elements did not differ significantly between sources. No significant overall correlation was found between essay quality and feedback quality for either source.

Al-Obaydi and Pikhart (2025) compared the impact of peer and AI assessments on 16 EFL students equally divided across two universities, one in Iraq and one in the Czech Republic. Participants (aged 22-23; gender-

balanced) were randomly assigned to 2 groups – peer assessment (n=8), and AI-based assessment (n=8). There were 4 participants from each university in each group. All participants produced an essay on a new topic each week for four weeks. The topic was uniform was all participants. Researchers developed a five-category rubric (main idea/focus, organisation and format, language use and style, originality, and creativity) and trained peer assessment group participants in its use. Each student in this group would use the rubric to give feedback to a different peer (from their own university) every week. Meanwhile AI-assessment group feedback participants used their personal ChatGPT accounts to first ask ChatGPT if it could revise their essays "based on main idea/focus, organisation and format, language use and style, originality, and creativity" (p.12). Once ChatGPT answered that it could, the students would upload their essay, ask ChatGPT to mention its flaws, and then rate their work out of 15. The quantitative data collected were the scores across the writing tasks. While students worked on their essays, researchers observed classes, recorded student behaviours, and gathered qualitative insights through evaluation sheets. These insights were related to the effectiveness, quality/perception of feedback, as well as feedback's impact on learning outcomes and engagement. While both groups improved, results indicated that the peer assessment group has higher final writing scores [7–14/15 (Iraq), 8–14/15 (Czech)] when compared to the AI feedback group [6–12/15 (Iraq & Czech)]. An analysis of the researchers' observation notes revealed that the peer group had benefitted more emotionally and cognitively from reciprocal interactions and engagement. Meanwhile, the AI feedback group received quicker, more accurate, and comprehensive feedback which was less emotionally engaging.

The study's findings about the difference in the quality of AI and peer feedback are suggestive and worthy of further investigation; however, a number of limitations must be mentioned. Limited sample size aside, it is unclear why the study did not use the same rubric for both feedback groups to ensure greater consistency in measuring the effectiveness of feedback across two mediums rather than two conditions (with and without rubrics). Also, while Group A wrote their essays by hand in class, Group B typed theirs at home and printed them. This meant that both sets of essays were produced in different conditions, which may explain the difference in scores more than the evaluative dispositions of human raters and AI raters. Finally, in the absence of a common rater (such as a teacher) across groups, one cannot speculate on which group benefitted more

2.4 Summary and research gaps

A review of the literature shows that written corrective feedback is generally beneficial for improving student writing proficiency, with most students preferring and benefitting the most from teacher feedback (Lv et al., 2021; Maas, 2017; Yang et al., 2006). However, teachers are often unable to give detailed corrective feedback

due to constraints of time and class size (Hattie & Timperley, 2007). In such cases, peer and AI feedback both serve as acceptable supplements to limited teacher feedback (Double et al., 2020; Latifi & Noroozi, 2021; Wang & Han, 2022; Zhang, 2020; Huawei & Aryadoust, 2023). Nevertheless, there are limitations to both. Peer feedback is often distrusted by students, especially in L2 contexts (Adams et al., 2011; Philp et al., 2010). It has also been found that peer feedback often skews toward praise or surface-level comments (Cho et al., 2006; Tsivitanidou et al., 2011). Meanwhile, AI feedback excels at local error detection (grammar, mechanics) but struggles with global features (content, coherence) (McCurry, 2010; Li et al., 2015; Stevenson & Phakiti, 2019). Explicit rubrics and training increase reliability and bring peer ratings and comments closer to teacher quality, while better prompts and rubrics can do the same for AI feedback (Falchikov & Goldfinch, 2000; Schunn et al., 2016; García-Varela et al., 2025).

Even though a growing number of studies have examined either peer feedback (e.g., Cho & McArthur, 2010; Schunn et al., 2016) or automated writing evaluation (e.g., Ranalli & Yamashita, 2022; Yu & Xie, 2025), very few have compared the two directly. Those that do (e.g., Banihashem et al., 2024; Al-Obaydi & Pikhart, 2025) report mixed results and often suffer from methodological inconsistencies, such as using different rubrics across groups or providing feedback sequentially rather than under controlled conditions. This leaves open the question of how AI and peer feedback compare when held to the same evaluative standards, especially if this comparison is across specific writing domains. Therefore, this study seeks to compare AI and peer feedback under such conditions, within a EFL setting higher education in Karachi, Pakistan, which presents a different context compared to Western and East Asian universities which have primarily been the site of previous studies.

Accordingly, this study asks the following research questions:

> RQ 1 Does the type of feedback received (AI, peer, or none) significantly affect participants' overall improvement, as well as their domain-specific gains in content, language use, and mechanics across two drafts, while controlling for initial language proficiency?
>
> RQ 2. Do AI and peer reviewers assign significantly different overall or domain-specific (Content, Language Use, Mechanics) scores to participants' first drafts when using the same rubric?
>
> RQ 3. How do AI- and peer-generated feedback differ in participants' perceptions and attributions and independently coded quality characteristics?

Chapter 3: Research Methodology

3.1 Research Setting

This research was conducted at a public-sector women's college in Karachi, Pakistan. The institution offers a four-year undergraduate program that prepares students to teach elementary grades (1–8). Women from all over the city attend the college. The college uses Microsoft Teams as its official learning management system (LMS). Every student is provided with a laptop, which they are required to bring to class. There are around 200 female students enrolled at the college at a time, with each cohort averaging between 50-60 students. Most students are recent Higher Secondary School graduates in their late teens or early-to-mid twenties, while some are in their thirties and early forties.

As Pakistan's largest metropolitan city, Karachi is home to many ethno-linguistic communities. Urdu is the first language of a majority of the city's population (Pakistan Bureau of Statistics, 2023). However, English, as a matter of policy, is supposed to be the Medium of Instruction (MoI) for higher secondary education and is the co-official language of the country. In practice, the use of English in most low-to-medium cost schools (whether public or private) remains limited (Ashraf et al., 2021; National Educational Policy, 2024).

Students at this teacher-education college are explicitly trained to give effective feedback. Pedagogical courses emphasize the "sandwich approach," in which critique is bracketed by praise. In addition, students at the college are required to read a policy document on effective feedback—covering comprehensibility, specificity, respectfulness, and actionability—at least once per term before completing faculty evaluation forms.

The college has also made a concerted effort to embrace generative AI. The college's Academic Integrity Policy states that students are allowed to use AI chatbots for their academic work as long as they are transparent about AI use and take full responsibility for the accuracy of their submitted work. The college's Academic Writing Centre runs mandatory workshops where students are trained to use tools (ChatGPT, Bard, CoPilot, etc) for text and image generation. Some assignment briefs explicitly require students, as future teachers, to use AI chatbots effectively for tasks such as lesson planning. Nevertheless, some writing assignments in Year 1 forbid students from using AI, and students are aware that a number of homebased assignments have been converted to in-person tests to mitigate overreliance on AI.

3.2 Research Aim and Questions

AI chatbots can affect teaching and learning both positively and negatively. Their use is ubiquitous; according to UNESCO (2025), nearly two-thirds of secondary school students globally are using AI to produce schoolwork.

On one hand, students' overreliance on AI can lead to a degradation of their core academic and cognitive skills, such as critical or creative thinking. Instantaneous responses by AI chatbots may increase student impatience and make students less likely to persevere with challenging tasks which are important for essential for experiential learning (Zhai et al., 2024). Conversely, AI has a democratising effect on educational attainment. AI chatbots help English-language learners communicate comprehensibly and become proficient by providing instant and personalized language support and real-time corrections (Pang, 2025). Additionally, AI can aid teachers in creating lesson plans, classroom resources, and providing detailed feedback to students (Paiva & Bittencourt, 2020; Burner & Wærness, 2025). Also, students often enjoy AI-integrated lessons (Kosmas et al., 2020).

AI feedback offers exciting possibilities for second language education, particularly in contexts where large class sizes limit the efficacy of teacher feedback and peers often lack the linguistic proficiency to provide actionable comments (Roshan et al, 2022; Allen & Mills, 2014). Situated within a higher-education EFL context where students are already familiar with generative AI, this research compares the quality and effectiveness of feedback provided by AI with that given by peers. In doing so, the study contributes to the growing body of literature on AI–peer feedback comparisons and offers insights into its potential in an understudied context such as Pakistan. The overarching aim is to explore the effectiveness and potential of AI-generated feedback in supporting student learning. To address this aim, the study poses three main research questions.

> RQ 1 Does the type of feedback received (AI, peer, or none) significantly affect participants' overall improvement, as well as their domain-specific gains in content, language use, and mechanics across two drafts, while controlling for initial language proficiency?
>
> RQ 2. Do AI and peer reviewers assign significantly different overall or domain-specific (Content, Language Use, Mechanics) scores to participants' first drafts when using the same rubric?
>
> RQ 3. How do AI- and peer-generated feedback differ in participants' perceptions and attributions and independently coded quality characteristics?

The first research question examines whether feedback type differentially contributes to measurable improvements in students' writing quality between drafts, both overall and within specific domains, independent of participants' English language proficiency. Writing quality is operationalised as rubric-based scores on both drafts of participants' essays, capturing overall quality as well as performance in content, language use, and mechanics.

The second research question examines whether reviewer type (AI or peer) resulted in significantly different scores being assigned to participants' first drafts. Each student's draft was rated once — either by an AI chatbot or by a peer reviewer — depending on their group allocation. Both AI and peer reviewers were given the same prompt (Appendix E) and rubric (Appendix D), which measured writing quality at two levels: an overall (global) score and three domain-specific scores (content, language use, and mechanics). The analysis compares whether the scores produced by AI differ significantly from those produced by peers.

The third research question investigates differences between AI- and peer-generated feedback from two perspectives. First, it examines participants' perceptions and attributions of the feedback they received, including whether they believed it was generated by AI or a peer and how they evaluated its usefulness. Second, it considers independent coding of the feedback itself, based on established quality criteria (see Kerman et al., 2022). Taken together, this question explores both participants' perceptions of feedback and the characteristics of AI- and peer-generated feedback as identified through coding against established quality criteria.

3.3 Design

The study followed a quasi-experimental design with three conditions: AI feedback, peer feedback, and control (no feedback). A call for participation was circulated in the college through posters and email announcements, and students signed up voluntarily. Initially, 45 participants signed up.

In an Introductory Session with the participants, the researcher (who is a faculty member at the college) explained the core aims and outlined the full procedure of the study. The participants were told that they would write an initial draft, give feedback to a peer, then be assigned to a group condition, and revise their final draft in light of the feedback they receive (or, in the case on the control condition, their own reflection). The participants were told that the submission of the final draft would be followed by a survey where they would answer questions about the feedback they received. They were also informed that they would not be explicitly told what feedback condition they had been assigned to until the study was completed. The researcher provided the participants with an information sheet (Appendix B) and consent form (Appendix A).

In the second meeting with the researcher, the participants returned the signed consent forms and took a 30-to-40-minute CEFR English test on their phones. This was the student-version of Core Skills tests by the British Council's mobile app 'English Score'. The researcher recorded the students' numerical scores, CEFR-based level, and ages. These details were given to the research assistant (RA). For pseudonymisation purposes, the RA assigned each participant a unique identification code, which was not shared with the researcher at the time.

The RA (who is also a faculty at the college) added the participants to an MS Teams class. The researcher was not given access to this Teams class. The participants then attended Session 1 with the RA. Session 1 was named so because it was when the students wrote their first drafts. The researcher was not present in this session.

The aforementioned MS Teams class contained the writing rubric (Appendix D) and feedback template (Appendix E) that participants would need in order to give feedback. First, the RA gave each participant a slip of paper with her assigned code. The participants were asked not to share this information with anyone else. Next, the participants were reminded that they would be writing an essay. The RA told the participants that their essay would be evaluated on a writing rubric. The RA gave all participants 10-minutes to read through the rubric. After this, the RA used the 'restrict access' feature on Teams to ensure that participants could no longer view the rubric.

Then, each participant was asked to download a file titled 'yourcode_essay' from MS Teams. The file contained the following prompt.

> Some people think school should start later in the day so students can sleep more.
> Do you agree or disagree?
> Explain your opinion and give **at least two reasons and examples**.
> Write **300-350 words**. You have **40 minutes**.

The RA gave the participant 40 minutes to write the first draft of their essay. The participants were told that no identifying information should be found in the draft, which should contain their assigned code, but not the names. When they were done writing, the participants saved to a shared folder on the Teams class. The participants titled the theircode_essay. For example, a student assigned the code MS321 would save their essay as MS321_essay. While the goal of these codes was to maintain anonymity, it quickly became clear to the RA that the Teams interface displayed participants names as 'uploaded by' information.

To assign reviewers to essays, all the paper slips containing codes were replaced in a bowl which was passed around. Participants randomly picked codes from a bowl, and this decided whose work they would review. Once again, the participants were given viewing access to the rubric. The were asked to read through the essays assigned to them carefully, evaluate them in light of the rubric, and give feedback and scores based on feedback template which was also provided to them on Teams. The RA instructed the participants to put their feedback comments and scores after the essay text, and to avoid in-line comments or annotations. Since the names of the essays' authors had been revealed, the RA ensured that the participants do not talk to each other so that they could not to reveal the authorship of the essay they reviewed. Nevertheless, every participant did know whose essay she was reviewing. Research has found that giving feedback to peers can help students improve their own drafts (Lundstrom & Baker, 2009). Each participant was required to give feedback to a peer so that any benefits of the feedback-giving process were held constant across the groups.

The participants were given 20 minutes to give feedback. After they had done this, the participants were no longer able to access the essays until the second group writing session. Following the peer feedback activity, the RA used an online randomiser to allocate participants to the three conditions: AI Feedback, Peer Feedback, Control (No Feedback). For the Peer group, the RA made no changes to the feedback.

For the AI group, the RA replaced peer comments with feedback generated by ChatGPT (GPT-4o). The ChatGPT feedback was generated using the participants' first drafts and the same rubric (Appendix) and feedback prompt (Appendix) which were given to peer reviewers as inputs. The RA made an entirely new ChatGPT account to generate feedback so that the chatbot would not draw on previous chats to influence the type of feedback produced.

For the Control group, the RA removed any peer comments and added a prompt which said 'Please read the above carefully and think about how it can be improved. In your second draft, make these improvements.' The RA also assigned entirely new codes to each essay, given that the older ones could be matched to names by the participants. The researcher was not involved in these stages.

Around this time, end-of-term examinations at the college had begun. Due to participants' differing examination schedules, the second draft was written in small groups rather than in a single, large group session. Session 2, therefore, was therefore staggered over the course of two weeks, with different groups of students meeting the research assistant (RA) at different times. Attrition during this period led to uneven final group sizes: AI feedback (n = 12), peer feedback (n = 11), and control (n = 13). 36 participants completed the study.

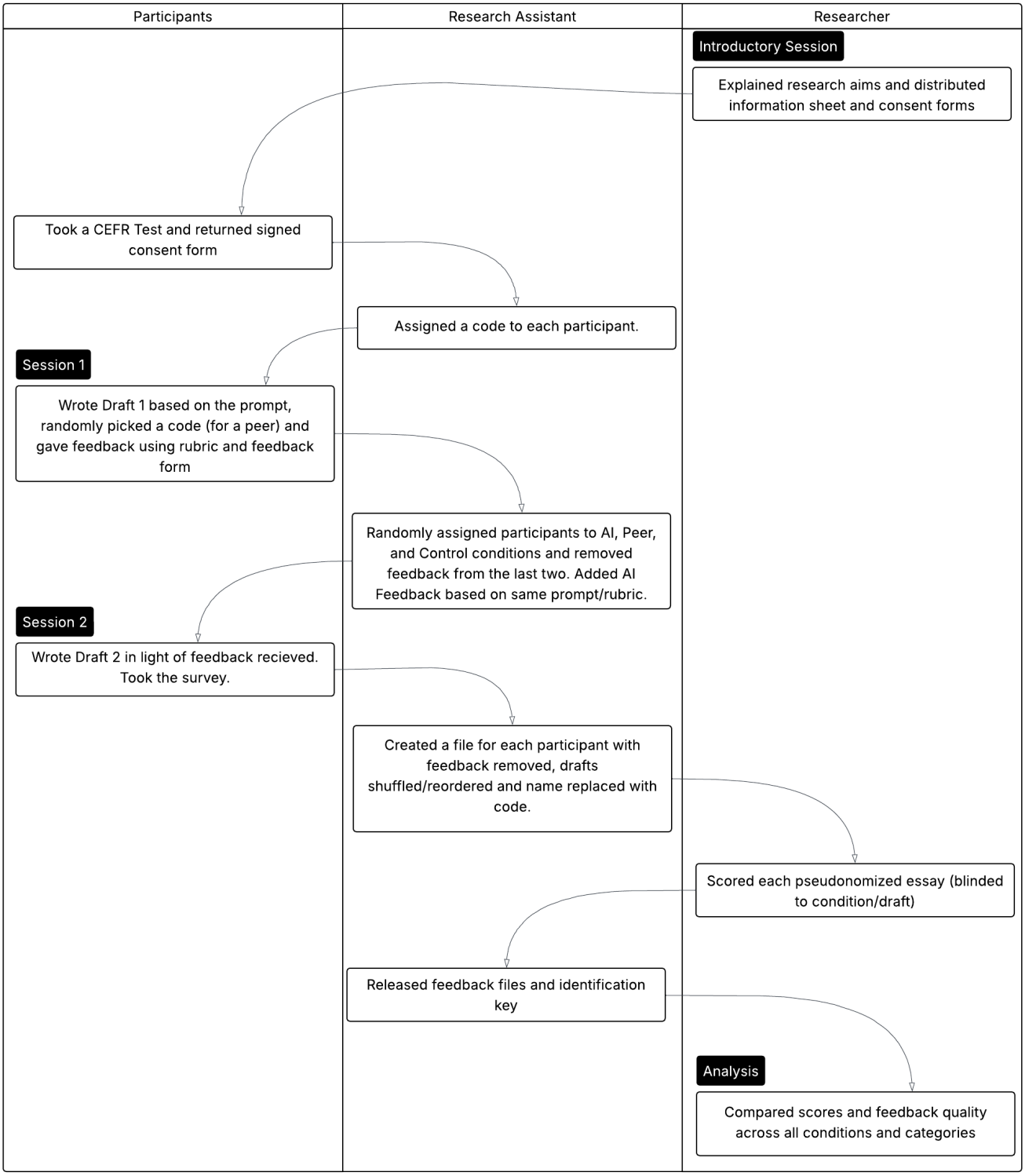
For each small group, the RA allowed the participants to view their individual first draft together with their assigned feedback only at the moment their revision session began. This ensured that, although participants began working on their second drafts at different times, each participant received feedback immediately before commencing their minute revision period. Maintaining this timing was important to control for the potential influence of feedback immediacy on its uptake and effectivenes Participants had 40 minutes to revise their work in light of the feedback provided (or, for the control group, based on their own reflection). Crucially, participants were not told whether the feedback they received was from peers or AI. They did know that AI feedback would have replaced peer feedback for some of them.

Participants began by reviewing the feedback given to them, copying their first draft beneath the feedback in the same document and then making changes in the copied text based on the feedback (or their own reflection, in the control group) to produce a second draft. Upon completing their second draft, participants completed a brief survey (Appendix F) designed to capture their perceptions of the revision process and the feedback they received or gave. Items asked about the perceived usefulness of revising, whether and from whom they believed they had received feedback, their reasons for this judgement, the perceived impact of giving feedback to others, and their preferences for types of support to improve writing.

Then, the RA collected all revised drafts, removed all feedback, and shuffled the order of drafts so that, for each participant, either the first or second draft appeared first. These pseudonymised drafts were provided to the researcher, who, blinded to feedback condition, participant, and draft order, scored both drafts of each essay using the 54-point rubric. The two drafts for a given participant were in the same file, and so scored consecutively by the researcher, who did not know which was Draft 1 and which was Draft 2.

After scoring was complete, the RA revealed key details to the researcher, allowing the researcher to compile scores by participant, draft, and condition. This quantitative data was analysed to answer the first two research questions. Once this was done, the researcher was then given access to the feedback messages which has been produced by AI and peers. The researcher could now see each student essay with either AI, peer, or no feedback. The researcher was also able to access to the survey responses. Both of these data sources were used to conduct a qualitative analysis of feedback characteristics and student perception in response to RQ3. Figure 1 outlines this process diagrammatically.

Figure 1: Research Process Flowchart

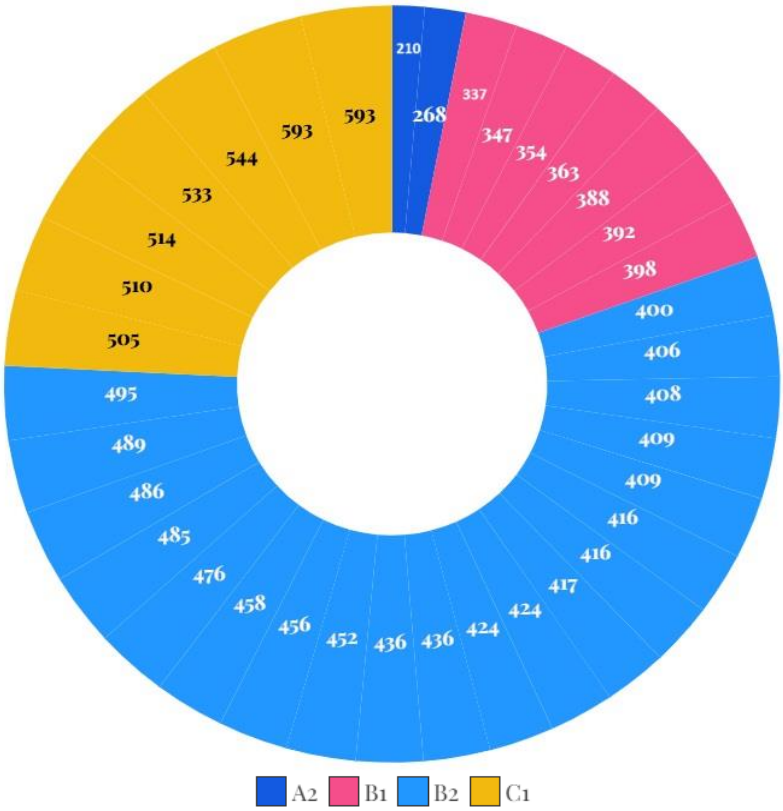


3.4 Participants

36 female undergraduate students participated in this study. All participants were enrolled in a four-year B.Ed (Hons) programme for elementary (Grades 1 – 8) teachers. The participants' ages ranged from 19 – 40 years, with an average age of 26.1 years. Students from across all four years participated in the project: Year 1 (n=10), Year 2 (n=3), Year 3 (n=14), and Year 4 (n=9).

The participants undertook a 30-to-40-minute 'Core Skills' online test of English grammar and comprehension (reading and listening) using the British Council's EnglishScore app. The results revealed that the participants had an average English proficiency score of around 435 on the app. This corresponds to the CEFR B2 level. According to the British Council (2025), a B2-level English learner can "use an extensive variety of phrases and complex grammatical structures to present detailed descriptions on subjects related to [their] fields of interest" (p. 30). The lowest participant score was 210 (A2) and the highest score was 593 (C1 level). Figure 2 shows all participant scores, grouped by CEFR level (see British Council (2025) for a detailed description of each CEFR level alongside its score range).

Figure 2: Parcitipants' Proficiency Scores and CEFR Levels



Note. Each segment represents an individual participant's score on the Core Skills placement test. Scores are arranged by CEFR level (A2, B1, B2, C1) and displayed in ascending order within each level.

3.5 Instruments

In order to respond to the core research questions, a mixed-methods approach was utilized. Quantitative and qualitative data were collected using four key instruments.

3.5.1 English Proficiency Test

Language proficiency was operationalised as the score received by each participant on the student version of the *Core Skills* test available on the British Council's EnglishScore App. The test was chosen for its credibility and practical advantages: developed by the British Council, a globally recognised authority in English language assessment, it can be administered entirely via a mobile phone and takes only 30–40 minutes to complete. Additionally, the test was already familiar to students at the college. Also, the test scores enable meaningful comparison of participants' proficiency against internationally recognised benchmarks as they are aligned with CEFR levels.

The Common European Framework of Reference for Languages (CEFR) is a global standard for describing language proficiency. Developed by the Council of Europe, it is widely used for language teaching, testing, and curriculum design worldwide, with six levels from A1 to C2. EnglishScore measured four proficiency levels: A2 – C1.

3.5.2 Writing Rubric

The rubric was based on the publicly-available Education Northwest (2021)' Six Traits of Writing framework. (see also: Kozlow & Bellamy, 2004). It was shortened considering the limited word count of the essay and edited based on the researcher's experience of teaching English to students at the college. The edited rubric assessed writing performance across three dimensions: Content and Organisation, Language Use, and Mechanics. Each dimension contained specific criteria, and every criterion was rated on a six-point scale. The rubric is reproduced in full in Appendix D.

The feedback form (Appendix E) given to students and ChatGPT (GPT-4o) was based the writing domains identified in this rubric, alongside general sentence starters to prompt both positive and critical feedback.

3.5.3 Survey

The student survey (Appendix F) was designed in Microsoft Forms to capture participants' perceptions of the feedback they received and their experiences with the revision process. It included both closed-ended and open-ended items. Closed-ended questions used Likert-type scales (e.g., *"To what extent did revising your first draft help you improve your essay?"* with responses ranging from *Not at all* to *Extremely*) and multiple-choice items (e.g., identifying the perceived source of feedback and preferred modes of support for

improving writing). Open-ended questions invited students to explain their reasoning (e.g., *"What makes you think the feedback you got was from a peer or AI?"*) or reflect on their learning (e.g., *"How did giving feedback to another student help you improve your essay?"*). Together, the items elicited both quantitative and qualitative data on participants' perceptions of AI- and peer-generated feedback.

3.5.4 Coding Scheme

Kerman et al. (2022)'s coding was used to code feedback responses into five categories: Affective (positive or negative emotion [praise, sympathy, etc.]); Cognitive–Description (summary of essay); Cognitive–Identification (pinpoint problem); Cognitive–Justification (explain why); Constructive (actionable recommendations. This scheme was selected because it is comprehensive, grounded in a well-established theoretical foundation, and has since been applied to compare AI- and peer-generated feedback (see: Banihashem et al., 2024).

3.6 Ethical Considerations

This project and its instruments were reviewed and received ethics clearance from the Central University Research Ethics Committee of the University of Oxford (see Appendix C). In addition, approval was also granted by the ethics committee of the college that served as the site of data collection. Participants were provided with a detailed participant information sheet (Appendix B) and consent form (Appendix A), and the study was explained to them in full. They were informed that participation was entirely voluntary. To protect anonymity, participants are not referred to by name but instead assigned codes (P1, P2, P3, ...). All data were either stored electronically in a secure server or kept in a locked cabinet, treated as confidential, and accessible only to the researcher and research assistant.

Chapter 4: Results

This chapter presents the results of the study, organised by research question. The first two questions are addressed through quantitative data analysis. A total of 36 participants took part in the study. Given this modest sample size, inferential results should be interpreted with caution. The final research question incorporates both quantitative and qualitative components.

RQ 1 Does the type of feedback received (AI, peer, or none) significantly affect participants' overall improvement, as well as their domain-specific gains in content, language use, and mechanics across two drafts, while controlling for initial language proficiency?

A one-way ANOVA showed no significant differences in baseline CEFR proficiency scores among the three groups, $F(2, 32) = 0.22$, $p = .801$. This indicates that participants in the AI, Peer, and Control groups did not differ in initial language proficiency.

1.1 Overall improvement

This sub-question asks whether the type of feedback received (AI, peer, or none) significantly impacts participants' overall performance across two drafts after controlling for their language proficiency. To answer, first, the gain in overall scores was calculated by subtracting the Draft 1 scores from the Draft 2 scores across all three feedback conditions (AI, peer, no feedback):

$$\text{Gain\_Total} = \text{Draft 2} - \text{Draft 1}$$

We needed to compare the mean gain scores of the AI feedback (n=12), peer feedback (n=11), and control (n=13) groups, while accounting for the influence of the participants' initial language proficiency. For the gain scores, Levene's test confirmed homoskedasticity ($p = .721$) across groups, but a Shapiro–Wilk test indicated non-normality for the AI group ($p = .034$). To account for this a 1000-samples bootstrapped one-way ANCOVA with proficiency test score as a co-variate was run. Linearity was confirmed ($R^2=0.027$), and homogeneity of regression slopes was indicated by a non-significant relationship between language proficiency score and Feedback Condition, $F (3, 32) = .505$, $p = .682$.

Although the AI Feedback group showed the highest mean gain ($M = 5.00$, $SD = 5.27$), followed by the Peer Feedback group ($M = 3.91$, $SD = 5.03$) and the Control group ($M = 2.92$, $SD = 4.50$), the results of the ANCOVA indicated that there was no significant effect of feedback condition on total gain scores after

adjusting for English language proficiency, F (2, 32) = 0.55, p = .58, η^2 = .033. The co-variate was not a significant predictor of gain, F (2, 32) = 0.55, p = .58, η^2 = .033.

Table 2: Descriptive Statistics for Total Gain Across Three Feedback Conditions

Condition	Total Gain			
	n	M	SD	95% bootstrapped BCa CI
Group 1: AI Feedback	12	5.00	5.27	[2.53, 8.00]
Group 2: Peer Feedback	11	3.91	5.03	[1.30, 7.32]
Group 3: Control (No Feedback)	13	2.92	4.50	[.60, 5.38]

Note. *M*= mean; *SD*= standard deviation; confidence intervals are based on 1000 bias-corrected and accelerated bootstrap samples

Table 3: ANCOVA Summary

Source	df	F	p	Partial η^2
Proficiency Score (covariate)	1, 32	0.92	.35	.03
Condition	2, 32	0.55	.58	.03
Error	32	—	—	—

Note. N = 36. Gain_Total = Draft 2 – Draft 1.

1.2 Improvement in Writing Domains (Content, Language Use, Mechanics)

This sub-question asks whether the type of feedback received (AI, peer, or none) significantly impacts participants' performance in writing subcategories (Content, Language Use, Mechanics) across two drafts after controlling for their English language proficiency.

First, the difference in scores across the two drafts in each of the three sub-categories was calculated for participants in each of the three feedback groups (AI, peer, control):

$Gain_{cont}$ = Draft 2 scores for Content and Organisation – Draft 1 scores for Content and Organisation

$Gain_{lang}$ = Draft 2 scores for Language Use – Draft 1 scores for Language Use

$Gain_{mech}$ = Draft 2 scores for Mechanics – Draft 1 scores for Mechanics

$Gain_{cont}$ scores were normally distributed across AI (p=.100), Peer (p=.277), and Control (p=.319) conditions, and the data was homoscedastic [F (2, 33) = 2.158, p = .132]. However, $Gain_{lang}$ scores were normally distributed across the AI (p=.667), but not the Peer (p=.003) or Control conditions (p=.002) and

the data was heteroskedastic [F (2, 33) = 4.211, p = .024]. Finally, $Gain_{mech}$ scores were not normally distributed for any feedback condition ($p < .05$ for all), though the data was homoscedastic [F(2, 33) = 0.637, p = .535]. Within each group, the relationship between the covariate (English proficiency) and the outcome (gain score) was linear.

$Gain_{mech}$ scores also violated the assumption of the homogeneity of regression slopes, F (2, 30) = 3.988, p = .029. As Johnson (2016) notes, such a violation indicates that the effect of the covariate differs across groups, rendering an ANCOVA inappropriate for this outcome. A violation of the homogeneity of regression slopes assumption cannot be remedied by bootstrapping, which addresses only distributional problems such as non-normality or heteroskedasticity, but not structural problems in the model specification. Therefore, to ensure interpretive comparability across all three rubric subcategories, a non-parametric test was used to answer RQ 1.2. The results of the resultant Krushkal-Wallis H Test are summarized in Table 4

Table 4: Results for Gain Scores in Writing Categories by Feedback Condition

Category	Group	Mean Rank	H (df = 2)	p-value
$Gain_{cont}$	AI	16.54	0.688	.709
	Peer	19.95		
	Control	19.08		
$Gain_{lang}$	AI	23.63	4.977	.083
	Peer	17.41		
	Control	14.69		
$Gain_{mech}$	AI	21.88	4.532	.104
	Peer	20.18		
	Control	13.96		

The AI feedback condition had the highest mean ranks for Language Use and Mechanics. In contrast, for Content and Organization, the Peer Feedback group had the highest mean rank. A series of Kruskal–Wallis H tests revealed no statistically significant differences in gain scores between the three feedback conditions (AI, peer, none) for Content and Organisation [H(2) = .688, p = .709], Language Use [H(2) = 4.977, p = .083], and Mechanics [H(2) = 4.532, p = .104].

A key part of RQ 1.2 is that participants initial language proficiency be accounted for. Since the Kruskal–Wallis H test cannot accommodate covariates, a series of linear regressions were conducted within each feedback condition (AI, Peer, Control) to examine whether language proficiency predicted gains in performance.

Table 5: Linear Regressions For Gain Scores by Proficiency Score Within Each Feedback Condition

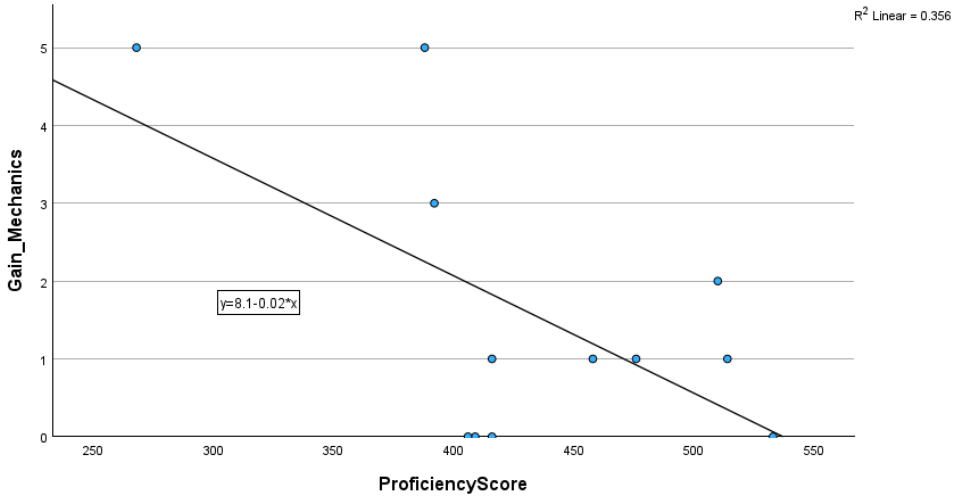
Feedback Condition		$Gain_{cont}$			$Gain_{lang}$			$Gain_{mech}$		
	N	R^2	β	p	R^2	B	p	R^2	β	p
Group 1: AI	12	.214	-.462	.130	.104	-.322	.307	.356	-.597	.041
Group 2: Peer	11	.022	.148	.663	.003	.054	.874	.001	-.034	.922
Group 3: Control	13	.059	-.243	.424	.002	0.48	.876	.151	.389	.189

Note. N=sample size per group. R^2 = variance in gain score explained by proficiency. β = standardized regression coefficient.

1.2.1 AI Feedback Group (N=12)

Participants who received AI feedback did not show significant gains in scores for the writing subcategories of Content and Organisation (p=.130) or Language Use (p=.307). However, the AI feedback group did show a significant (p=.041) moderate-to-strong (R^2 = .356) negative link between in the between scores for the Mechanics subcategory and language proficiency. In other words, higher proficiency participants seem to have made fewer gains in mechanics after receiving AI feedback compared to their lower proficiency peers. Figure 3 visually illustrates this relationship.

Figure 3: Scatterplot of proficiency scores and gains in mechanics in the AI feedback condition.



1.2.2 Peer Feedback Group (N=11)

In the Peer feedback group, participants initial language proficiency did not significantly predict gains in Content and Organisation (p=.663), Language Use (p=.874), or Mechanics (p=.922). The small R^2 values (.002, .003, .001 for the three subcategories respectively) further indicate that almost none of the differences in gains was explained by participants' English-language proficiency.

1.2.3 Control (No Feedback) Group (N=13)

Participants received no feedback in the control group. Instead, they were asked to make changes to their first draft based on their own proofreading and review. In this group, participants initial language proficiency did not significantly predict gains in Content and Organisation (p=.424), Language Use (p=.876), or Mechanics (p=.186).

RQ 2. Do AI and peer reviewers assign significantly different overall or domain-specific (Content, Language Use, Mechanics) scores to participants' first drafts when using the same rubric?

Participants received either peer or AI scores on the first draft of their essays. Different sets of essays were scored by peers and ChatGPT using the same rubric (Appendix D). This question asks if there was a difference in the scores assigned by ChatGPT or Peers either on a holistic or category-specific level.

2.1 Overall scores

For the overall (total) score, the Shapiro–Wilk test indicated that the assumption of normality was met for both the Peer Feedback group (p = .836) and the AI Feedback group (p = .381). An independent samples t-test was used to compare the scores assigned by ChatGPT (GPT-4o) and the scores assigned by peer reviewers. The data met the assumption of homogeneity of variance (p=.871) and normal distribution for the Peer (p=.836) and AI (p=.381) raters. The first draft mean score was 33.64 (SD = 4.6) for the Peer Feedback group and 33.58 (SD = 4.1) for the AI Feedback group. No significant difference was found between AI- and peer-assigned scores, t(21) = .029, p = .977.

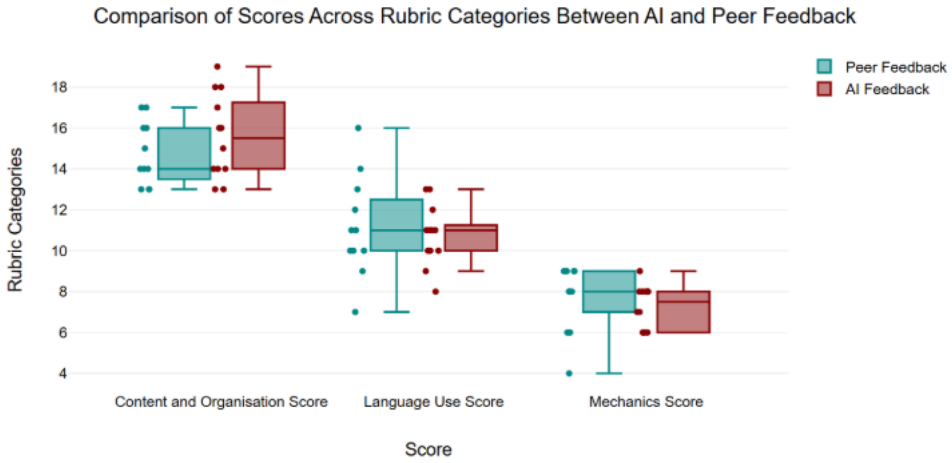
2.2 Writing Subcategory (Content and Organisation, Language Use, Mechanics) Scores

Tests of normality were conducted using the Shapiro–Wilk test for each feedback condition (AI vs. Peer) across the three rubric subcategories.

Results indicated that the assumption of normality was met for Content and Organisation (Peer: p = .090; AI: p = .304) and Language Use (Peer: p = .886; AI: p = .570). However, the assumption of normality was violated for Mechanics (Peer: p = .005; AI: p = .032), where both groups showed significant departures from

a normal distribution. Even with the violation of assumption of normality for this last group, a box plot (Figure 4) of AI- and Peer-assigned scores across all three writing rubric categories indicated roughly similar shapes and spreads for AI and Peer groups—satisfying a key assumption for using Mann–Whitney U tests.

Figure 4: Distribution of AI and Peer Scores Across Content, Language Use, and Mechanics.



A Mann-Whitney U test was used to compare the scores assigned by Peer raters with those assigned by ChatGPT (GPT-4o) for the writing subcategories (Content and Organisation, Language Use, Mechanics). The first draft mean scores and standard deviations for the scores assigned by the two raters across each subgroup in given in Table 7. No significant differences were found for Content and Organisation ($U = 50$, $p = .316$), Language Use ($U = 61.5$, $p = .778$), or Mechanics ($U = 45$, $p = .179$). While peer reviewers tended to assign higher scores for Mechanics (Mean Rank = 13.91) compared to AI (Mean Rank = 10.25), this difference was not statistically significant.

Table 6: Descriptive Statistics for Scores Assigned by Category Across Peer and AI Feedback Conditions

Category	Group	Mean Rank	Mann-Whitney U	p (Asymp. Sig) (2-tailed)
Content and Organisation	AI	13.33	50	.316
	Peer	10.55		
Language Use	AI	11.63	61.5	.778
	Peer	12.41		
Mechanics	AI	10.25	45	.179
	Peer	13.91		

RQ 3. How do AI- and peer-generated feedback differ in participants' perceptions and attributions and independently coded quality characteristics?

Following the completion of their second draft, participants responded to a survey where they were asked to identify if the feedback given to them was by a Peer or an AI agent and justify their choice. These survey responses were thematically coded to see how participants perceive AI vs Peer feedback. Furthermore, the feedback messages generated by AI or Peer agents were coded against an established typology of feedback messages to locate any differences between the two.

RQ 3.1. Can participants correctly identify whether feedback was given by AI or a peer, and what linguistic/stylistic cues do they cite?

All 36 participants completed a survey (Appendix F) which asked them to judge if they received feedback, and whether that feedback was from peers or AI. 32 participants identified the source of feedback correctly. 1 participant was unsure. 3 participants judged incorrectly. Of these, 1 control group participant thought the message she had received to edit her draft herself was from an AI agent, another thought it was from a peer. 1 Peer feedback group participant thought her feedback was AI-generated.

Participants were also asked 'What makes you think the feedback you got was from a peer or AI?' After removing Control group participants and any incorrect identifications, the remaining 32 open-ended responses were coded in NVivo and four recurring categories were identified: Accuracy or Relevance, Comprehensiveness, Specificity, and Tone. These categories were not imposed a priori but emerged through repeated coding and comparison, representing the main ways participants differentiated between AI- and peer-generated feedback. A detailed explanation of these categories, alongside illustrative quotes from participants, is given below.

3.1.1 Accuracy or Relevance

Participants frequently judged feedback by how relevant and correct it appeared. The correctness included judgements about the feedback agent's advice as well as its own language proficiency. Peer comments were sometimes perceived as inaccurately worded or misaligned with the essay (e.g., "The feedback I found was human written because it was irrelevant" [Peer, P18]; "has small mistakes because it comes from a student teacher like me" [Peer, P22]; "not very relevant. I recieved 5 in punctuation but feedback says that I need to use more punctuations." [Peer, P13]. In contrast, AI was often praised for its accuracy ("the level of english also shows that it is AI" [AI, P3], "constructive" [AI, P4]).

3.1.2 Comprehensiveness

Participants usually described AI feedback as detailed and exhaustive, while peer comments were sometimes viewed as brief or limited. For example: "The feedback is not much in detail" [Peer, P14] versus "Because it was very detailed" [AI, P3] or "It was from AI, as it has more details with many mistakes [identified]" [AI, P11].

3.1.3 Specificity

AI feedback was recognised by its ability to pick up on minute and particular errors (e.g., "pointing out specific sentences" [AI, P2], "highlighting minor areas" [AI, P3], "As it highlighted each and every point where it needed improvement. So, I guess it was an AI tool." [AI, P1], "[gave] specific topic in brackets" [AI, P6]). In contrast, Peer feedback was not typically described as providing this kind of specific, sentence-level commentary with specific errors identified.

3.1.4 Tone

Tone served as another distinguishing factor. Peer feedback was often characterised as more natural and personable: "I got feedback from my peer because Peer feedback is friendly" [Peer, P22], "use of frank language" [Peer, P20], "the language was natural" [Peer, P21]. One the other hand, one comment indicated that AI feedback was perceived as more neutral or professional: "It was following the constructive tone" [AI, P4].

RQ 3.2. How do the quality characteristics of AI vs. peer feedback differ when coded using an established typology of feedback messages.

ChatGPT (GPT-4o) and peers used the same rubric (Appendix D) and feedback template (Appendix E) to produce 12 and 11 units of feedback, respectively. To analyse the nature and quality of feedback across the two conditions (AI and Peer), I employed the feedback typology developed by Kerman et al. (2022).

The typology comprises five categories of feedback: Affective (A), which includes positive or negative emotions such as praise, encouragement, or criticism; Cognitive–Description (CD), a summary of the essay content; Cognitive–Identification (CI), pinpointing and localising specific problems in the text; Cognitive–Justification (CJ), providing reasons or elaborations to explain why an issue is problematic; and Constructive (Con), offering recommendations for revision or improvement. AI and Peer Feedback responses were coded in accordance with this typology. An example of the type of comments each category included are given in Table 7.

Table 7: Feedback Excerpt Examples Coded According to the Selected Typology

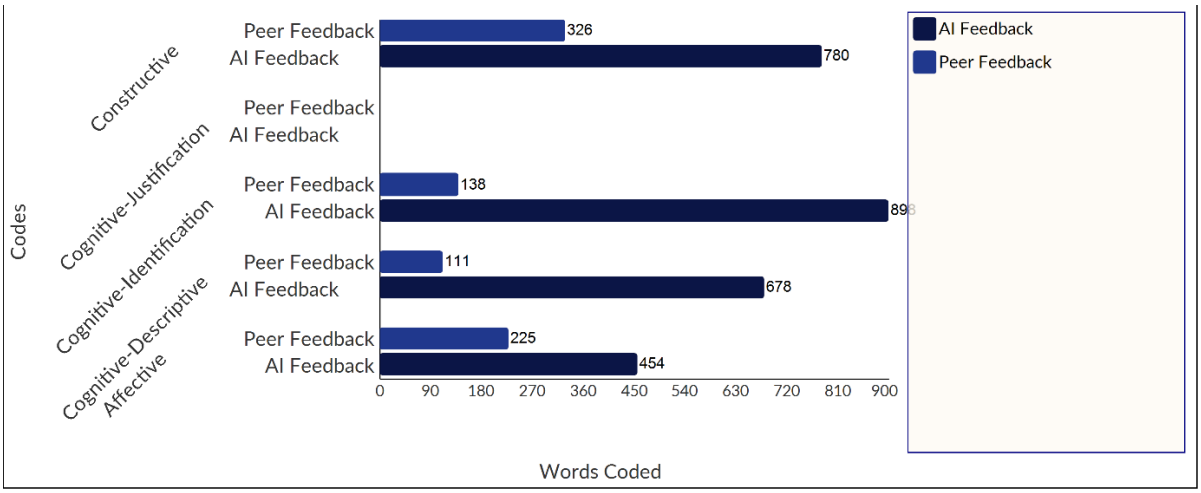
Feedback Excerpts	
AI Feedback	**Peer Feedback**
Strengths: I think the writer did a great job clearly stating their position in the thesis. The essay consistently argues that schools should start early, emphasizing the benefits like discipline, time management, and improved memory. The use of personal reflections, like the morning being good for health and habit-building, effectively supports the thesis. **Areas for Improvement:** However, the essay could be even stronger if it included more concrete examples and specific details to back up each point. For instance, the section on time management could benefit from a specific student example to enhance relevance and connection. Additionally, the essay lacks a distinct conclusion that ties all the points together with a sense of closure, making the argument feel somewhat abrupt.	**Content and Organization:** I think the writer did a great job because she started with a relatable issue going on, and highlighted two reasons in her thesis statement that aligned with her beliefs. Her points were clear and stated well, backed up with examples. However, her writing could be even stronger if she had built on the same points and given one detailed example instead of touching upon multiple examples.

Key	A		CD		CI		CJ		Con		

As Table 7 shows, sentence starters such as 'I think the writer did a good job' or 'However, the essay could be even stronger' were not coded when seemed either verbatim or very similar to the sentence starters given in the feedback form (Appendix E) itself, as these phrases were the result of using a preset structure for the feedback and not natural AI or Peer productions.

The AI and Peer Feedback outputs were coded and analysed in NVivo. Figure 5 offers a summary of results in the form of the word count for each code for each feedback condition.

Figure 5: Word Count for Each Code by Feedback Condition



The results show that AI tended to give more feedback across all domains. However, this difference was most pronounced in the 'Cognitive-Identification' domain. This was followed by a large difference in the Constructive, and the Cognitive-Descriptive domains. Neither agent gave feedback that could be interpreted as being justification for produced errors. This may be due to the restrictions of the feedback template and rubric.

To answer RQ 3.2, the Peer and AI Feedback outputs were analysed according to the Kerman et al. (2022)'s typology. This thematic analysis showed that AI and Peer agents produced both similar and different messages in within the five feedback domains. The analysis below highlights these supported by illustrative quotes.

3.2.1 Affective

In the Affective domain, both ChatGPT (GPT-4o) and Peers were likely to comment on the clarity of ideas presented ("clearly stating their position in the thesis" [AI, P1], "establishing a clear thesis statement" [AI, P2], "clearly stating" [AI, P5], "she clearly mention what is the essay is about" [Peer, P13], "points were clear" [Peer, P16]). Neither feedback agent offered explicitly negative comments about a writer's work. Both were likely to couch their criticism within advice for improvement.

While AI gave more affective feedback, it also tended to be more formulaic in its responses, often mentioning how the presentation of ideas was 'effective' or 'consistent' or language use 'appropriate' [for e.g., P1, P2, P3, P6, P12].

Meanwhile, peer responses were less predictable in their responses, but also less likely to use specific evaluative adjectives in their praise, sometimes simply stating that the writer "did good job" or "provide

good content" [Peer, P19] or was good at "using vocabulary" [Peer, P15]. Sometimes, peers praised each other more effusively (e.g., "excellent essay with good info" "awesome" [Peer, P20]). Overall, ChatGPT was more likely to pinpoint specific aspects of the essay and quote them when discussing them, will Peer feedback was more generalised and often vague.

While a majority of both Peer-praise and AI-praise focused on the global aspects of the essays (e.g., "readable to the general audience" [Peer, P16], ChatGPT was more likely to offer Mechanics related comments (e.g., "capitalization is mostly accurate" [AI, P12]).

3.2.2 Cognitive

3.2.2.1 Cognitive-Descriptive

ChatGPT vastly outproduced peers when it came to Cognitive-Descriptive comments. The AI review offered succinct summaries of essays. ChatGPT-summaries mostly focused on both global, content language concerns (e.g., "with multiple reasons from historical, scientific, religious, and personal perspectives" [AI, P8]; "with a strong focus on why schools should start early — primarily for better health, brain efficiency, and academic success" [AI, P10]; "using formal academic language and varying sentence structures" [AI, P7]). Sometimes, the writer's mechanical choices would be summarised (e.g., "mostly correct grammar and spelling throughout the essay. Commas and periods are generally well used" [AI, P10]).

Peers did not offer much Cognitive-Descriptive Feedback. The comments that did appear were mostly focused on content-level summaries (e.g. "why there should be no school in the morning" [Peer, P18]), and only one of these restated details (e.g., "a separate point in each paragraph such as habit, health, time management" [Peer, P14]). In terms of language use, two peers (P14, P22) highlighted the writers' effective use of transition words, while ChatGPT offered no comments on transition words.

3.2.2.2 Cognitive-Identification

In the Cognitive-Identification subdomain, ChatGPT's feedback was precise and diagnostic. It often pointed to specific sentences or paragraphs where problems of repetition, coherence, or logic occurred, or would reproduce full words and phrases from the text. For example, it highlighted redundancies directly ("the second paragraph repeats the same point about habit formation without introducing new insights" [AI, P3]) or drew attention to particular, problematic phrases ("for example, "peace their minds" is unclear" [AI, P5]; "(e.g., article usage: "the afternoon schools" not "afternoon schools")" [AI, P9].

Likewise, peers also pointed to specific issues in the text, often with examples. When commenting on the redundancy of an essay, a peer pointed out "For example: Many people believe that school should start later in the day. Some believe that if school started later the students get more time to sleep. These two are in the same first paragraph" (P14). While peers did not produce direct quotations as often as ChatGPT, they did point to the specific areas where errors had occurred ("concluding paragraph, this seems vague and brought to a abrupt halt" [Peer, 16]; "conclusion can be enhanced because it seemed very obvious" [Peer, P20]). Also, some peers have offered in-text annotations and referred back to them in the end comments ("see underlined in 2nd last paragraph" [Peer, P15], "because in some place unnecessary articles were used that word pointed out" [Peer, P16]).

Notably, peers offered much less feedback in this subdomain when compared to AI. Peer comments were often shorter and less likely to reproduce examples. Very few peer reviewers focused on mechanical concerns. Mostly, either peer reviewers did not explicitly identify any errors (P17, P18, P19, P22, P23, P20), or identified only content and language level errors (P20, P21). In contrast, ChatGPT pointed to instances of mechanical errors in every single feedback message either generally (e.g., "A few comma splices and run-ons (long sentences that should be broken or better connected" [AI, P12]; "commas are often missing or misused" [AI, P9]) or specifically ("a few minor grammar errors, such as "our minds works10 times better" and "the key of a healthy, wealthy and a successful life" [AI, P2].)

3.2.2.3 Cognitive-Justification

Neither peers nor ChatGPT offered comments in the Cognitive-Justification subdomain. This may be because the feedback template specifically asked reviewers to first offer specific praise and then specific improvement recommendations.

3.2.3 Constructive

In terms of Constructive comments. ChatGPT consistently offered more detailed and specific feedback. Its advice focused on specific aspects of improvement for both global aspects of writing (e.g., it asked writers to develop arguments with "concrete examples", vary "sentence length and structure", and tighten conclusions by "reaffirming the main position" [AI, P1; P3; P7] and local aspects (e.g., it asked writer's to correct "subject-verb agreement" and employ "formal transitions between ideas" [AI, P6; P9]. When offering corrections for mechanical or language use issues, ChatGPT would often include concrete examples or rephrased structures. For example, after recommending "more formal transitions," ChatGPT offered model transitions ("e.g., "Firstly," "Secondly," "Furthermore""), and, after pointing out that "peace their minds" is "unclear", it recommended the alternative phrase "help students feel calm and focused" [AI, P9; P5]. This

was a recurring them in AI-generated feedback (e.g., “schools should starts → schools should start” [AI, P6]; "in deserted regions of Africa → “in the desert regions of Africa” [AI, P7]; “"roved" should be "proved"” [AI, P8]).

In contrast, while peers did also point to specific and general content, language, and structural improvements (e.g., elaborating “a little more in the concluding paragraph” [Peer, P16], adding a “clear conclusion” and “transition words” [P18], “adding academic research or data” [P20] and employing “proper grammar/usage” [P21]), they rarely suggested how to do this concretely. Advice to “work on her grammar” [P20], “use more punctuations” [P13], or have “proper grammar/usage” [P22] is generalised. Some peers provided more targeted suggestions, such as offering the proverb “Early to bed and early to rise...” [Peer, P14] or recommending breaking long sentences into smaller ones [Peer, P15]. However, these were exceptions rather than the norm.

Notably, although AI feedback was more structured and comprehensive, it was also more predictable. Firstly, ChatGPT followed the feedback template more staunchly and began advice with ‘The essay would be stronger if...’ followed by examples. Secondly, ChatGPT seemed to give more consistent advice. The suggestion to add ‘more concrete examples’ (for instance) was present in nearly every AI feedback message. Meanwhile, patterns such as these did not appear in peer feedback.

Summary of Results

Table 9 presents a summary of the results of each research question alongside the various ways in which data were analysed.

Table 8: Summary of Results

RQ		Analysis	Result
RQ1 Does the type of feedback received (AI, peer, or none) significantly affect participants’ overall improvement, as well as their domain-specific gains in	Overall	One-way ANCOVA comparing overall gain scores across feedback conditions with proficiency test score as a covariate.	Mean Gain: AI (5.00) > Peer (3.91) > Control (2.92). No significant effect of feedback condition on total gain score, $F(2, 32) = 0.55$, $p = .58$, $\eta^2 = .033$. English language proficiency was not a significant predictor of gain, $F(2, 32) = 0.55$, $p = .58$, $\eta^2 = .033$.

content, language use, and mechanics across two drafts, while controlling for initial language proficiency?	Domain-specific	A series of Kruskal–Wallis H tests comparing gain scores in Content and Organisation, Language Use, and Mechanics across the three feedback conditions.	<u>Content and Organisation</u> Mean Rank: Peer (19.95) > Control (19.08) > AI (16.54). Difference = ns. <u>Language Use</u> Mean Rank: AI (23.63) > Peer (17.41) > Control (14.69). Difference = ns. <u>Mechanics</u> Mean Rank: AI (21.88) > Peer (20.18) > Control (13.69). Difference = ns
		A series of within-group linear regressions testing whether language proficiency predicted gain scores in each feedback condition.	<u>AI Feedback Group</u> Content/Organisation: ns Language Use: ns Mechanics: significant negative relationship, $p = .041$, $R^2 = .356$ (higher proficiency → fewer gains). <u>Peer Feedback Group</u> All domains: ns <u>Control Group</u> All domains: ns
RQ2 Do AI and peer reviewers assign significantly different overall or domain-specific (Content, Language Use, Mechanics) scores to participants' first drafts when using the same rubric?	Overall	Independent samples t-test comparing scores assigned by ChatGPT and peer reviewers	Difference between AI- and peer-assigned scores was not significant, $t(21) = .029$, $p = .977$.
	Domain specific	Mann-Whitney U test to compare Peer-rater scores with ChatGPT scores for writing domains (Content and Organisation, Language Use, Mechanics)	No significant differences for Content and Organisation, Language Use, or Mechanics.

RQ3. How do AI- and peer-generated feedback differ in participants' perceptions and attributions and independently coded quality characteristics?	Participants	Thematic analysis of participants' responses to survey question (how did they know who gave them feedback?)	Accuracy/Relevance: AI often described as accurate and aligned; peer comments sometimes perceived as irrelevant or containing errors. Comprehensiveness: AI feedback viewed as more detailed; peer feedback described as brief or limited. Specificity: AI recognised for identifying sentence-level and minor errors; peer feedback seen as less precise. Tone: Peer feedback characterised as friendly/natural; AI once described as neutral or professional.
	Feedback Typology	Feedback messages coded according to Kerman et al. (2022)'s typology	Constructive: AI (780 words) > Peer (326). AI offered more specific, example-based revision advice; Peer feedback was briefer and often general. Cognitive–Identification: AI (678) > Peer (111). AI consistently pinpointed sentence-level issues, including mechanics; Peers mostly identified content/language issues with fewer examples. Cognitive–Descriptive: AI (454) > Peer (225). AI produced fuller essay summaries; Peer summaries were fewer and mostly content-level. Affective: AI (398) > Peer (138). Both praised clarity and thesis; AI more formulaic, Peer more variable and effusive. Cognitive–Justification: None in either condition (likely due to template constraints).

Chapter 5: Discussion and Conclusion

This study compares the effects of AI- and peer-generated feedback on undergraduate EFL students' writing improvement in Karachi, Pakistan, focusing on the gain in participants' scores across two drafts. It also investigates the differences in feedback (in both scores and comments) between AI and peer raters when both used the same rubric and feedback template to assess works by writers of similar language abilities. Finally, the study investigates students' perceptions of AI or peer feedback (see Table 9 for a summary of results). This section discusses the main insights gained from the study.

Effectiveness of Feedback by Condition and Learner Proficiency

Although participants who received AI feedback showed the largest overall mean improvement across drafts ($M = 5.00$), followed by those receiving peer feedback ($M = 3.91$) and no feedback ($M = 2.92$), the differences were not statistically significant ($p = .58$). Similarly, domain-specific comparisons revealed no significant advantage of one feedback source over another: peer feedback was associated with slightly higher gains in content and organisation, whereas AI feedback was linked to higher gains in language use and mechanics. These patterns suggest that feedback type may matter less than expected for short-term improvement, echoing meta-analytic findings that feedback effectiveness is highly conditional (Hattie, 2009; Kang & Han, 2015). This also aligns with Fang et al.'s (2025) earlier findings that there are no significant differences in the effectiveness of peer and AI feedback.

Moreover, participants' initial language proficiency did not significantly predict gains in content and organisation or language use across any of the feedback conditions. This finding diverges from Kang & Han (2015)'s earlier meta-analyses of 22 studies which report proficiency as a key moderator of feedback effectiveness, with beginners benefiting little and advanced learners showing the strongest gains. One explanation for this discrepancy may be the relatively narrow proficiency band of participants in this study, with most clustered around the B2 level (see Figure 2). In such a context, proficiency may exert less influence than other factors, such as students' engagement with the revision process or the clarity of feedback received. Another possible explanation is the short timescale: feedback effects on higher-order aspects of writing, such as organisation and idea development, may require more sustained practice across multiple writing tasks to manifest significantly.

Importantly, the only significant regression effect emerged in the AI group, where higher proficiency students made fewer gains in mechanics ($p = .041$, $R^2 = .356$). This counterintuitive finding suggests that AI feedback may be most beneficial for learners at beginner or intermediate proficiency, who still make frequent

surface-level errors. More advanced learners may find AI corrections less useful if they have fewer mechanical errors. This aligns with broader findings suggesting that AI feedback is most effective when well-matched to learners' current needs and proficiency levels (Zhang et al., 2024). Unlike Yan and Zhang's (2024) study, which found that higher-proficiency and more digitally competent students benefitted more from ChatGPT feedback because of their stronger ability to craft effective prompts and interpret responses, the present study positioned the research assistant as the intermediary for generating AI feedback. By removing prompt crafting from learners' control, the design may have reduced the influence of proficiency-linked differences in digital competence.

Taken together, these findings suggest that while feedback type did not significantly shape short-term gains, AI feedback may hold particular value for intermediate learners at the surface level.

Reliability of AI vs Peer as Evaluators

The differences between the overall and domain-specific (Content and Organisation, Language Use, Mechanics) scores between ChatGPT and student reviewers were non-significant. The convergence between AI and peer scores may partly be attributed to the use of a detailed rubric (Appendix D), which reduced ambiguity in rating criteria, and a feedback form (Appendix F), which structured reviewers' responses. By constraining both human and AI raters to the same evaluative framework, the study design may have amplified score reliability while minimising differences between reviewer types.

This supports earlier findings that detailed rubrics can enhance the reliability of peer assessments (Falchikov & Goldfinch, 2000; Schunn et al., 2016) and also aligns with recent work showing that structured prompts can increase the consistency of AI-generated ratings (García-Varela et al., 2025). The present study adds to this growing evidence by showing that rubrics not only aid individual reviewer reliability but may also help align human and AI raters with one another (Yu & Xie, 2025).

Yet this convergence may reflect the constraining effect of the rubric and feedback template, which standardised how both AI and peers evaluated writing. While such structure amplified reliability, might it also have reduced variation in feedback style and depth, thereby contributing to the lack of significant differences in participants' gain scores between the two feedback conditions? To explore this possibility, the next section turns to a qualitative analysis of feedback comments and participants' perceptions.

Perceptions and Characteristics of Feedback

Nearly all participants in the AI and peer feedback groups were able to accurately distinguish between AI and peer feedback. Notably, their perceptions mostly aligned with the independent coding and analysis of the feedback performed using Kerman et al. (2022)'s typology.

Participants perceived AI feedback as more accurate, comprehensive, and effective at identifying specific, sentence-level errors when compared to peer feedback, echoing earlier studies showing that AI offers more sentence rewrites and metalinguistic information on errors when compared to human reviewers (Lin & Crosthwaite, 2024; Yu & Xie, 2025). In this study, an analysis of the feedback messages revealed the same, with AI producing more extensive summaries of student work, and pinpointing more errors while offering corrections. The difference between AI and peers was substantive. This is dissimilar to Banihashem et al. (2024) who coded AI and peer feedback messages using the same typology and found that students were more likely to pinpoint errors. A possible reason for this discrepancy may be differences in student proficiency. While most students in this study were at a CEFR B2 level of proficiency, Banihashem et al. (2024) did not report on their participants' overall language proficiency. Previous studies have shown that higher-proficiency L2 students provide more detailed feedback when compared to their lower proficiency peers (Allen & Mills, 2016).. Given that explicit solutions are an effective component of WCF (Wu & Schunn, 2021; Nelson & Schunn, 2009), AI chatbots seem to be important feedback tools.

Even though the feedback template gave equal weight to all writing domains, many students either ignored the mechanics part of the template, or gave very limited comments. In contrast, ChatGPT consistenly followed the template and offered more comments. While aligned with previous studies in that peer feedback tends to be limited (Cho et al., 2006) and that AI offers more mechanical corrections when compared to human raters (Lin & Crosthwaite, 2024), it nevertheless did not find that AI stuggled to give more holsitic/global-level feedback on content and organisation (McCurry, 2010; Stevenson & Phakiti, 2019, Wang, 2020; Lang et al., 2019). One reason for this may be the use of a feedback template and rubric, which explictly asked ChatGPT to give feedback on global concerns.

Participants often noted inaccuracies and misalignment in peer feedback, but also characterized it by its friendly and natural tone. An analysis of feedback messages found that while AI offered a greater volume of praise, its praise was likely to be formulaic and consistent ("clearly stating" [AI, P5]) while peers varied greatly in the type of praise they offered – sometimes slilted (“did good job” [Peer, P19]) and other times effusive (“awesome” [Peer, P20]) Relatedly, previous studies have noted that peer tendency to provide irrelevant or overly positive remarks (Cho et al., 2006; Tsivitanidou et al., 2011).

Finally, neither AI nor peer reviewers offered comments that could be reasonably seen as justifying why errors occured. This is somewhat suprising, especially given that Banihashem et al. (2024) found a number of such comments in their study. Perhaps, however, this speaks one of the key limitations of using a detailed rubric and template. While such a process does lead to a greater consistency among raters, it may also constrain the type of comments they are likely to produce. Practitioners must weigh the costs and benefits of feedback templates; in this study, while Cognitive-Justification comments were likely not produced due to the sentence structures and limits in the feedback template, AI and peer raters did, neverthless, display some consistency in scoring.

Limitations and Recommendations

Here, several design features temper the claims but also indicate clear directions for improvement. First of all, the studies small, single-site sample and two-draft window limited statistical power and external validity; multi-site studies with larger cohorts and delayed post-tests would better capture durable learning and transfer. Secondly, the studies' participants clustered around CEFR B2, narrowing proficiency variance; future work should recruit across bands (B1–C1) and stratify randomisation to test proficiency as a moderator more sensitively. Also, the shared rubric and feedback template likely equalised raters but constrained comment types (e.g., the absence of cognitive-justification comments), foregrounding a reliability–validity trade-off; subsequent studies could trial lighter templates or prompts that explicitly elicit "why" rationales and dialogic follow-ups, since the present study's more prescriptive template appeared to constrain reviewers and may partly explain the absence of justification comments.

Furthermore, AI feedback was generated by a research assistant using fixed prompts and templates: this standardised delivery but reduced ecological validity relative to learner-prompted use. Replications should compare mediated vs. student-prompted AI, logging prompts and interactions to examine how digital competence shapes outcomes.

In this study, peer identities were inferable during the peer feedback stage. While this mirrors authentic classrooms and may enhance accountability, it can induce social-desirability or expectancy effects. A useful extension is to experimentally vary anonymity (anonymous vs. known peers) to estimate its impact on comment quality and perceptions.

Finally, the outcomes of this study were limited to rubric-based gains on a single argumentative writing task using one AI model and configuration. Future research could expand the scope by including additional genres of writing, comparing different AI models or versions, and experimenting with alternative prompting scaffolds. Incorporating a teacher-feedback condition or hybrid approaches (e.g., AI→peer or peer→AI

sequences) would also help test the complementarity of these feedback sources and provide stronger evidence on whether AI and peers can sustainably supplement teacher feedback while easing teachers' workload.

Conclusion

This study offers an early, contextually grounded comparison of AI- and peer-generated feedback in a Pakistani EFL teacher-education setting. Quantitatively, feedback type did not yield significant short-term gains, although AI showed a small advantage for mechanics, and both AI and peers converged in scoring when guided by a common rubric and template. Qualitatively, participants perceived—and coding corroborated—AI feedback as more accurate, comprehensive, and specific, while peer feedback was friendlier but less systematic; both seldom provided justification, likely reflecting template constraints. Taken together, the results suggest that structure (rubrics/prompts) can equalise evaluators, and that AI may be best leveraged for local accuracy while peers contribute relational support. Practically, this indicates that AI and peer review can serve as viable supplementary feedback sources in classrooms to complement teacher feedback and help lighten teacher workload—provided appropriate scaffolds (clear rubrics, feedback templates, and brief training) are in place. Larger and longer-term studies can test these patterns across genres and proficiency bands and evaluate whether reliability gains translate into durable learning.

References

Adams, R., Nuevo, A., & Egi, T. (2011). Explicit and implicit feedback, modified output, and SLA: Does explicit and implicit feedback promote learning and learner-learner interactions? *Modern Language Journal, 95*(1), 42–63 https://doi.org/10.1111/j.1540-4781.2011.01242.x

Al-Jarrah, R. S. (2016). A suggested model of corrective feedback provision. *Ampersand*, *3*, 98–107. https://doi.org/10.1016/j.amper.2016.06.003

Allen, R., Benhenda, A., Jerrim, J., & Sims, S. (2021). New evidence on teachers' working hours in England. An empirical analysis of four datasets. *Research Papers in Education*, *36*(6), 657–681. https://doi.org/10.1080/02671522.2020.1736616

Allen, D., & Katayama, A. (2016). Relative second language proficiency and the giving and receiving of written peer feedback. *System*, *56*, 96–106. https://doi.org/10.1016/j.system.2015.12.002

Allen, D., & Mills, A. (2016). The impact of second language proficiency in dyadic peer feedback. *Language Teaching Research*, *20*(4), 498-513.

Apriani, E., Cardoso, L., Obaid, A. J., Muthmainnah, N., Wijayanti, E., Esmianti, F., & Supardan, D. (2024). Impact of AI-Powered ChatBots on EFL Students' Writing Skills, Self-Efficacy, and Self-Regulation: A Mixed-Methods Study. *Impact of AI-Powered ChatBots on EFL Students' Writing Skills, Self-Efficacy, and Self-Regulation: A Mixed-Methods Study Global Educational Research Review*, *1*(2), 57–72. https://doi.org/10.71380/gerr-08-2024-8

Ashraf, M. A., Turner, D. A., & Laar, R. A. (2021). Multilingual Language Practices in Education in Pakistan: The Conflict Between Policy and Practice. *SAGE Open*, *11*(1). https://doi.org/10.1177/21582440211004140 (Original work published 2021)

Bangert-Drowns, R., & Kulik, C. (1991). The Instructional Effect of Feedback in Test-like Events. *Review of Educational Research, 61*, 213–238. Retrieved from http://rer.sagepub.com/content/61/2/213.short

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, *21*(1), 23. https://doi.org/10.1186/s41239-024-00455-4

Biber, D., Nekrasova, T., & Horn, B. (2011). The Effectiveness of Feedback for L1-English and L2-Writing Development: A Meta-Analysis. *ETS Research Report Series*, *2011*(1), i–99. https://doi.org/10.1002/j.2333-8504.2011.tb02241.x

Bitchener, J., & Knoch, U. (2010). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing, 19*(4), 207–217. https://doi.org/10.1016/j.jslw.2010.10.002

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

Bonilla López, M. (2021). An Updated Typology of Written Corrective Feedback: Resolving Terminology Issues. *Revista Educación*. https://doi.org/10.15517/revedu.v45i1.43289

Brown, D., Liu, Q., & Norouzian, R. (2023). Effectiveness of written corrective feedback in developing L2 accuracy: A Bayesian meta-analysis. *Language Teaching Research*, 13621688221147374.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.2307/1170684

Cárcamo, B. (2020). Classifying written corrective feedback for research and educational purposes: a typology proposal. *PROFILE Issues in Teachers Professional Development*, *22*(2), 211–222. https://doi.org/10.15446/profile.v22n2.79924

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on Writing: Typology and Perceived Helpfulness of Comments from Novice Peer Reviewers and Subject Matter Experts. *Written Communication*, *23*(3), 260-294. https://doi.org/10.1177/0741088306289261 (Original work published 2006)

Cho, K., Schunn, C.D., & Wilson, R.W. (2006). Validity and re liability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*(4), 891–901. doi:10.1037/0022- 0663.98.4.891

Coe, M., Hanita, M., Nishioka, V., & Smiley, R. (2011). An Investigation of the Impact of the 6+ 1 Trait Writing Model on Grade 5 Student Writing Achievement. Final Report. NCEE 2012-4010. *National Center for Education Evaluation and Regional Assistance*.

Comajoan-Colomé, L., & Salguero, T. (2024). Typologies of Written Corrective Feedback. In *The TESOL Encyclopedia of English Language Teaching* (pp. 1–8). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118784235.eelt1023.pub2

Crosthwaite, P., Ningrum, S., & Lee, I. (2022). Research trends in L2 written corrective feedback: A bibliometric analysis of three decades of Scopus-indexed research on L2 WCF. *Journal of Second Language Writing*, *58*, 100934. https://doi.org/10.1016/j.jslw.2022.100934

Cui, Y., Schunn, C. D., & Gai, X. (2021). Peer feedback and teacher feedback: a comparative study of revision effectiveness in writing instruction for EFL learners. Higher Education Research & Development, 41(6), 1838–1854. https://doi.org/10.1080/07294360.2021.1969541

Daumiller, M., & Meyer, J. (2025). Advancing feedback research in educational psychology: Insights into feedback processes and determinants of effectiveness. *Contemporary Educational Psychology*, 102390. https://doi.org/10.1016/j.cedpsych.2025.102390

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1).

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2019). The Impact of Peer Assessment on Academic Performance: A Meta-analysis of Control Group Studies. *Educational Psychology Review*, *32*(2), 481–509. https://doi.org/10.1007/s10648-019-09510-3

Education Endowment Foundation. (2021). *Teacher feedback to improve pupil learning: Guidance report*. https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/feedback-approaches

Ellis, R. (2009). A typology of written corrective feedback types. *ELT Journal, 63*(2), 97–107. https://doi.org/10.1093/elt/ccn023

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System, 36*(3), 353–371. https://doi.org/10.1016/j.system.2008.02.001

Ende J. (1983). Feedback in clinical medical education. *JAMA*, *250*(6), 777–781.

Eraut, M. (2007). Learning from other people in the workplace. *Oxford Review of Education*, *33*(4), 403–422. https://doi.org/10.1080/03054980701425706

Espasa, A., & Meneses, J. (2010). Analysing feedback processes in an online teaching and learning environment: An exploratory study. *Higher Education*, *59*(3), 277–292. https://doi.org/10.1007/s10734-009-9247-4

Evans, N. W., Hartshorn, K. J., & Strong-Krause, D. (2011). The efficacy of dynamic written corrective feedback for university-matriculated ESL learners. *System*, *39*(2), 229-239.

Fang, Y., Tan, Y., Zuo, C., & Boubaker, A. (2024). AI generated vs. peer feedback in ESL writing: Effects on writing skill, self-efficacy, and enjoyment. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.5115438

Fazio, L. L. (2001). The effect of corrections and commentaries on the journal writing accuracy of minority- and majority-language students. *Journal of Second Language Writing, 10*(4), 235–249. https://doi.org/10.1016/S1060-3743(01)00042-X

Ferreira, A., Moore, J. D. and Mellish, C. (2007). A study of feedback strategies in foreign language classrooms and tutorials with implications for intelligent computerassisted language learning systems. *International Journal of Artificial Intelligence in Education, 17*, pp. 389–422.

García-Varela, F., Nussbaum, M., Mendoza, M., Martínez-Troncoso, C., & Bekerman, Z. (2025). ChatGPT as a stable and fair tool for automated essay scoring. *Education Sciences*, *15*(8), 946. https://doi.org/10.3390/educsci15080946

Gentry, R., & Wallace-Nesler, V. (2014). *Fostering writing in today's classroom*. Teacher Created Materials.

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal, 115*(4), 523–547. https://doi.org/10.1086/681947

Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Huawei, S. & Aryadoust, V. (2023). A systematic review of automated writing evaluation systems. *Education and Information Technologies, 28*, 771–795. https://doi.org/10.1007/s10639-022-11200-7

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, *39*(2), 83–101. doi:10.1017/S0261444806003399

Johnson, T. R. (2016). Violation of the homogeneity of regression slopes assumption in ANCOVA for two-group pre-post designs: Tutorial on a modified Johnson-Neyman procedure. *The Quantitative Methods for Psychology*, *12*(3), 253–263. https://doi.org/10.20982/tqmp.12.3.p253

Kang, E., & Han, Z. (2015). The Efficacy of written corrective feedback in improving L2 Written Accuracy: A Meta-Analysis. *Modern Language Journal*, *99*(1), 1–18. https://doi.org/10.1111/modl.12189

Keith T. Z. (2003). Validity and automated essay scoring systems. In Shermis M. D., & Burstein J. C. (Eds.), *Automated Essay Scoring; A Cross—disciplinary Perspective*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 147-168.

Kerman, N. T., Noroozi, O., Banihashem, S. K., Karami, M., & Biemans, H. J. A. (2022). Online peer feedback patterns of success and failure in argumentative essay writing. *Interactive Learning Environments*, *32*(2), 614–626. https://doi.org/10.1080/10494820.2022.2093914

Kluger, A. N., & DeNisi, A. S. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–84.

Knoblock, J., & Drake, J. (2005). *The 6 Traits of Writing*.

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for Language Teaching and Learning. *RELC Journal*. https://doi.org/10.1177/00336882231162868

Kozlow, M., and Bellamy, P. (2004). Experimental study on the impact of the 6+1 Trait® writing model on student achievement in writing. Portland, OR: Northwest Regional Educational Laboratory.

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. Longman.

Kulhavy, R. W. (1977). Feedback in written instruction. Review of Educational Research, 47, 211–232.

Lang, F., Li, S., & Zhang, S. (2019). Research on Reliability and Validity of Mobile Networks-Based Automated Writing Evaluation. *International Journal of Mobile Computing and Multimedia Communications (IJMCMC), 10*(1), 18-31. https://doi.org/10.4018/IJMCMC.2019010102

Langove, S. A., & Khan, A. (2024). Automated grading and feedback systems: reducing teacher workload and improving student performance. *Deleted Journal*, *13*(4), 202–212. https://doi.org/10.62345/jads.2024.13.4.16

Latifi, S., & Noroozi, O. (2021). Supporting argumentative essay writing through an online supported peer-review script. *Innovations in Education and Teaching International*, *58*(5), 501–511. https://doi.org/10.1080/14703297.2021.1961097

Lin, S., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System*, *127*, 103529. https://doi.org/10.1016/j.system.2024.103529

López, M. B., Van Steendam, E., Speelman, D., & Buyse, K. (2018). The differential effects of comprehensive feedback forms in the second language writing class. *Language Learning*, *68*(3), 813–850. https://doi.org/10.1111/lang.12295

Lou, N. M., & Noels, K. A. (2020). "Does My Teacher Believe I Can Improve?": The Role of Meta-Lay Theories in ESL Learners' Mindsets and Need Satisfaction. *Frontiers in psychology*, *11*, 1417. https://doi.org/10.3389/fpsyg.2020.01417

Lu, Q., Yao, Y., & Zhu, X. (2023). The relationship between peer feedback features and revision sources mediated by feedback acceptance: The effect on undergraduate students' writing performance. *Assessing Writing*, *56*, 100725. https://doi.org/10.1016/j.asw.2023.100725

Lundstrom, K. & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. Journal of Second Language Writing, 18(1), 30-43.

Lv, X., Ren, W., & Xie, Y. (2021). The Effects of Online Feedback on ESL/EFL Writing: A Meta-Analysis. *The Asia-Pacific Education Researcher*, *30*(6), 643–653. https://doi.org/10.1007/s40299-021-00594-6

Maas, C. (2017). Receptivity to learner-driven feedback in EAP. *Elt Journal*, *71*(2), 127-140.

MacArthur, C. A. (2007). Evaluation and revision. In S. Graham, C. A. MacArthur, & J. Fitzgerald (Eds.), *Best practices in writing instruction* (pp. 141–162). Guilford Press.

McCurry, D. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? *Assessing Writing*, *15*(2), 118–129. https://doi.org/10.1016/j.asw.2010.04.002

McDonald, F. J., & Allen, D. (1962) An investigation of presentation, response and correction factors in programmed instruction. *Journal of Educational Research, 55*, 502-507.

Ministry of Federal Education & Professional Training. (2024). *National Education Policy Development Framework – 2024*. Government of Pakistan. https://pie.gov.pk/SiteImage/Publication/NEPDF%202024%20(17.12.2024).pdf

Molenaar, I. (2022). The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. *Computers and Education: Artificial Intelligence, 3*, 100070.

Narciss, S., & Huth, K. (2004). How to Design Informative Tutoring Feedback for Multimedia Learning. In Niegemann, H., Brunken, R., & Leutner, D. (Eds.), *Instructional Design for Multimedia Learning* (pp. 181-196). Munster: Waxmann.

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science*, *37*(4), 375–401. https://doi.org/10.1007/s11251-008-9053-x

Pakistan Bureau of Statistics. (2023). Population Census 2023. In *Pakistan Bureau of Statistics*. Government of Pakistan. Retrieved August 1, 2025, from https://www.pbs.gov.pk/sites/default/files/population/2023/material/sindh_insight.pdf

Panadero, E., Romero, M., & Strijbos, J.W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fair ness and comfort. *Studies in Educational Evaluation, 39*(4), 195–203. doi:10.1016/j.stueduc.2013.10.005

Pang, S., Nol, E., & Heng, K. (2025). Generative AI as a Personal Tutor for English Language Learning: A Review of Benefits and Concerns. *International Journal of Changes in Education*. https://doi.org/10.47852/bonviewIJCE52023724

Paris, B. (2017). The Influence of Language Proficiency on Student Response to Direct and Indirect Written Corrective Feedback (Master's thesis, University of Calgary, Calgary, Canada). Retrieved from https://prism.ucalgary.ca. doi:10.11575/PRISM/28211 http://hdl.handle.net/11023/3991 Downloaded from PRISM Repository, University of Calgary

Park, S., Weng, W. (2020) The relationship between ICT-related factors and student academic achievement and the moderating effect of country economic index across 39 countries: using multilevel structural equation modelling. *Educ Technol Soc 23*(3):1–15

Patchan, M. M., Charney, D., & Schunn, C. D. (2009). A validation study of students' end comments: Comparing comments by students, a writing instructor, and a content instructor. Journal of Writing Research, 1(2), 124-152. https://doi.org/10.17239/jowr-2009.01.02.2

Pearson, W. S. (2022). A typology of the characteristics of teachers' written feedback comments on second language writing. *Cogent Education*, *9*(1), 2024937. https://doi.org/10.1080/2331186X.2021.2024937

Philp, J., Walter, S., & Basturkmen, H. (2010). Peer interaction in the foreign language classroom: What factors foster a focus on form? *Language Awareness, 19*(4), 261–279. https://doi.org/10.1080/09658416.2010.516831

Potter, R., & Fuller, D. (2008). My New Teaching Partner? Using the Grammar Checker in Writing Instruction. *The English Journal*, *98*(1), 36–41. http://www.jstor.org/stable/40503205

Ranalli, J., & Yamashita, T. (2022). Automated written corrective feedback: Error-correction performance and timing of delivery. Language Learning & Technology, 26(1), 1–25. https://doi.org/10.64152/10125/73465

Ryan, T., & Henderson, M. (2017). Feeling feedback: students' emotional responses to educator feedback. *Assessment & Evaluation in Higher Education*, *43*(6), 880–892. https://doi.org/10.1080/02602938.2017.1416456

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119–144.

Schunn, C., Godley, A., & DeMartino, S. (2016). The Reliability and Validity of peer review of writing in High School AP English Classes. *Journal of Adolescent & Adult Literacy*, *60*(1), 13–23. https://doi.org/10.1002/jaal.525

Sheen, Y., Wright, D., & Moldawa, A. (2009). Differential effects of focused and unfocused written corrective feedback in an ESL context. *System, 37*(4), 556–569. https://doi.org/10.1016/j.system.2009.09.002

Shermis, M. D., & Burstein, J. (Eds.) (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.10.4324/9781410606860Search in Google Scholar

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, *36*(2), 187–209. doi:10.1017/S0958344023000265

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153-189. https://doi.org/10.3102/0034654307313795 (Original work published 2008)

Simmons, J. (2003). Responders are Taught, Not Born. *Journal of Adolescent & Adult Literacy*, *46*(8), 684–693. http://www.jstor.org/stable/40017173

Skinner, B. F. (1968). The technology of teaching. New York: Appleton-Century-Crofts.

Smith, J. K., & Lipnevich, A. A. (2018). Instructional Feedback: Analysis, Synthesis, and Extrapolation. In A. A. Lipnevich & J. K. Smith (Eds), *The Cambridge Handbook of Instructional Feedback* (pp. 591–603). Cambridge University Press. https://doi.org/10.1017/9781316832134.034

Spandel, V. (2012). *Creating writers: 6 traits, process, workshop, and literature*. Pearson Higher Ed.

Stevenson, M. and Phakiti, A. (2019). Automated feedback and second language writing. In Hyland, K. and Hyland, F. (Eds.) Feedback in Second Language Writing. Cambridge: Cambridge University Press, pp. 125–141.

Stevenson, M., & Phakiti, A. (2019). Automated feedback and second language writing. *Feedback in second language writing: Contexts and issues*, 125-142.

Storch, N., & Wigglesworth, G. (2010). LEARNERS' PROCESSING, UPTAKE, AND RETENTION OF CORRECTIVE FEEDBACK ON WRITING: Case Studies. *Studies in Second Language Acquisition*, *32*(2), 303–334. doi:10.1017/S0272263109990532

Thorndike, E. L. (1927). The law of effect. *The American Journal of Psychology, 39,* 212–222. https://doi.org/10.2307/1415413

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*(2), 327–369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x

Truscott, J. (1999). The case for "the case against grammar correction in L2 writing classes": A response to Ferris. *Journal of Second Language Writing, 8*(2), 111–122. https://doi.org/10.1016/S1060-3743(99)80124-6

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing, 16*(4), 255–272. https://doi.org/10.1016/j.jslw.2007.06.003

Truscott, J., & Hsu, A. Y. (2008). Error correction, revision, and learning. *Journal of Second Language Writing, 17*(4), 292–305. https://doi.org/10.1016/j.jslw.2008.05.003

Tsivitanidou, O.E., Zacharia, Z.C., & Hovardas, T. (2011). Investigating secondary school students' unmediated peer assessment skills. *Learning and Instruction, 21*(4), 506–519. doi:10.1016/j.learninstruc.2010.08.002

Van Beuningen, C., De Jong, N. H., & Kuiken, F. (2008). The effect of direct and indirect corrective feedback on L2 learners' written accuracy. *International Journal of Applied Linguistics, 156*(2), 279–296. https://doi.org/10.1075/itl.156.02bea

Van Beuningen, C., De Jong, N. H., & Kuiken, F. (2012). Evidence on the effectiveness of comprehensive error correction in second language writing. *Language Learning, 62*(1), 1–41. https://doi.org/10.1111/j.1467-9922.2011.00674.x

Van De Weghe, R. (2004). "Awesome, dude!": Responding helpfully to peer writing. *English Journal, 94*(1), 95–99. doi:10.2307/4128855

van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive Load Theory and Complex Learning: Recent Developments and Future Directions. *Educational Psychology Review, 17*(2), 147–177. https://doi.org/10.1007/s10648-005-3951-0

Van Steendam, E., Rijlaarsdam, G., Sercu, L., & Van den Bergh, H. (2010). The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL. Unravelling Peer Assessment, 20(4), 316–327. https://doi.org/10.1016/j.learninstruc.2009.08.009

Wang, Y. J., Shang, H. F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning*, *26*(3), 234-257.

Wang, Z. (2020). Computer-assisted EFL writing and evaluations based on artificial intelligence: a case from a college reading and writing course. Library Hi Tech, 40(1), 80-97. https://doi.org/10.1108/LHT-05-2020-0113

Wang, Z., & Han, F. (2022). The effects of teacher feedback and automated feedback on cognitive and psychological aspects of foreign language writing: A Mixed-Methods Research. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.909802

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies an International Journal*, *3*(1), 22–36. https://doi.org/10.1080/15544800701771580

Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay scoring: Current applications and future directions* (pp. 36–54). New York: Routledge.10.4324/9780203122761.ch3Search in Google Scholar

Wisniewski B, Zierer K and Hattie J (2020) The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. Front. Psychol. 10:3087. doi: 10.3389/fpsyg.2019.03087

Wolery, M., Ault, M. J., Doyle, P. M., & Gast, D. L. (1986). Comparison of instructional strategies: A literature review. https://files.eric.ed.gov/fulltext/ED345418.pdf

Wu, Y., & Schunn, C. D. (2021). The effects of providing and receiving peer feedback on writing performance and learning of secondary school students. *American Educational Research Journal, 58*(3), 492–526. https://doi.org/10.3102/0002831220945266

Yan D, Wang J (2022) Teaching data science to undergraduate translation trainees: pilot evaluation of a task-based course. Front Psychol 13:939689. https://doi. org/10.3389/fpsyg.2022.939689

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing*, *15*(3), 179–200. https://doi.org/10.1016/j.jslw.2006.09.004

Yu, H., & Xie, Q. (2025). Generative AI vs. Teachers: Feedback Quality, Feedback Uptake, and Revision. *Language Teaching Research Quarterly*, *47*, 113–137. https://doi.org/10.32038/ltrq.2025.47.07

Zeevy-Solovey, O. (2024). Comparing peer, ChatGPT, and teacher corrective feedback in EFL writing: Students' perceptions and preferences. *Technology in Language Teaching & Learning*, *6*(3), 1482. https://doi.org/10.29140/tltl.v6n3.1482

Zhai, C., Wibowo, S. & Li, L.D. (2004). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learn. Environ. 11*, 28. https://doi.org/10.1186/s40561-024-00316-7

Yvdal, A., & Bergström, O. (2024). Chatgpt-4's effectiveness in providing feedback on argumentative writing in higher education: A case study.

Zhang, S. (2021). Review of automated writing evaluation systems. *Journal of China Computer-Assisted Language Learning*, *1*(1), 170-176. https://doi.org/10.1515/jccall-2021-2007

Zhang, Z. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, *43*, 100439. https://doi.org/10.1016/j.asw.2019.100439

Zheng Y, Yu S (2018) Student engagement with teacher written corrective feedback in EFL writing: a case study of Chinese lower-proficiency students. Assess Writ 37:13–24. https://doi.org/10.1016/j.asw.2018.03.001

UNESCO. (2025). In high-income countries, more than two-thirds of secondary-school pupils are already using generative AI tools to produce schoolwork. *UNESCO*. Retrieved from the UNESCO International Day of Education press release. dianova.org+7UNESCO+7Gadget+7

Burner, T., Lindvig, Y., & Wærness, J. I. (2025). "We should not be like a dinosaur"—Using AI technologies to provide formative feedback to students. *Education Sciences, 15*(1), 58. https://doi.org/10.3390/educsci15010058 arXiv+12MDPI+12ResearchGate+12

Kosmas, P., Nisiforou, E. A., Kounnapi, E., Sophocleous, S., & Theophanous, G. (2025). Integrating artificial intelligence in literacy lessons for elementary classrooms: A co-design approach. *Educational Technology Research and Development.* Advance online publication. https://doi.org/10.1007/s11423-025-10492-z

Paiva, R., & Bittencourt, I. I. (2020). Helping Teachers Help Their Students: A Human-AI Hybrid Approach. *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I, 12163*, 448–459. https://doi.org/10.1007/978-3-030-52237-7_36

Kohnke, L., Zou, D., & Su, F. (2025). Exploring the potential of GenAI for personalised English teaching: Learners' experiences and perceptions. *Computers and Education: Artificial Intelligence*, *8*, 100371. https://doi.org/10.1016/j.caeai.2025.100371

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, *27*, 1–18. https://doi.org/10.1016/j.jslw.2014.10.004

Fleckenstein J, Liebenow LW and Meyer J (2023) Automated feedback and writing: a multi-level meta-analysis of effects on students' performance. Front. Artif. Intell. 6:1162454. doi: 10.3389/frai.2023.1162454

Yan, D., & Zhang, S. (2024). L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study. *Humanities and Social Sciences Communications*, *11*(1), 1086. https://doi.org/10.1057/s41599-024-03543-y

Zhang, X., Zhang, P., Shen, Y., Liu, M., Wang, Q., Gašević, D., & Fan, Y. (2024). A Systematic Literature Review of Empirical Research on Applying Generative Artificial Intelligence in Education. *Frontiers of Digital Education*, *1*(3), 223–245. https://doi.org/10.1007/s44366-024-0028-5

Allen, D., & Mills, A. (2015). *Peer Feedback in the Academic English Classroom: A Pilot Study*

Roshan, A., Gurbaz, M., & Rahmani, S. (2022). The effects of large classes on English language teaching. *Integrated Journal for Research in Arts and Humanities*, *2*(2), 38.

Kadek Nina Harnin, Ni Nyoman Padmadewi, Ni Luh Putu Eka Sulistia Dewi, & Ni Komang Arie Suwastini. (2022). TEACHERS' PERCEPTION AND PRACTICES ON GIVING FEEDBACK ON STUDENTS' WORK DURING ONLINE LEARNING. *JISAE: Journal of Indonesian Student Assessment and Evaluation*, *8*(1), 55–65. https://doi.org/10.21009/jisae.v8i1.26011

Yu, S., Zheng, Y., Jiang, L., Liu, C., & Xu, Y. (2021). "I even feel annoyed and angry": Teacher emotional experiences in giving feedback on student writing. *Assessing Writing*, *48*, 100528. https://doi.org/10.1016/j.asw.2021.100528

British Council. (2025). *EnglishScore: Core Skills*. https://www.englishscore.com/wp-content/uploads/2024/07/Test_prep.pdf

Appendices

Appendix A: Sample Consent Form

Consent to take part in 'AI or Peer Feedback: What Works Best in Improving Writing?'

Central University Research Ethics Committee (CUREC) approval reference: Education (Educ) DREC - 1012627

Purpose of Study: This study investigates what kind of feedback is more effective in helping students improve their writing skills – AI or peer?

	Please initial each box if you agree with the statement
I confirm that I have read and understand the information sheet dated April 2025 for the above research. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.	☐
I understand that my participation is voluntary and that I am free to withdraw at any point until 30th June 2025, without giving any reason.	☐
I understand who will have access to personal data provided, how the data will be stored and what will happen to the data at the end of the project.	☐
I understand that I will not be identifiable from any publications.	☐
Use of quotations: Please indicate your preference (select *one* option): a) I do not wish to be quoted. **or** b) I agree to the use of quotations in research outputs if I am not identifiable.	☐
I give permission for you to contact me again to clarify information.	☐

I understand how to raise a concern or make a complaint.	
I agree to take part.	
I agree that my personal contact details can be retained in a secure database so that the researchers can contact me about future studies.	YES / NO

_____	dd / mm / yyyy	_____
Name of participant	Date	Signature

_____	dd / mm / yyyy	_____
Name of person taking consent	Date[1]	Signature

[1] To be signed and dated in the presence of the participant. Once this has been signed by both parties the participant should receive a copy of the signed and dated participant consent form. The original signed and dated consent form should be kept with the project's main documents, which must be kept in a secure location.

Appendix B: Sample Information Sheet

AI or Peer Feedback: What Works Best in Improving Writing?

Central University Research Ethics Committee Approval Reference: Education (Educ) DREC - 1012627

Introductory paragraph

You are being invited to take part in a research project. Before you decide whether to take part, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether you wish to take part.

Why is this research being conducted?

Feedback is an important component of learning. This research is being conducted to investigate what kind of feedback is more effective in helping students improve their writing skills. The two types of feedback being compared are peer feedback and AI feedback.

Why have I been invited to take part?

We want to learn from the real experiences of students who receive feedback from peers and/or AI. Your participation can provide us with authentic information about which method of feedback is best and help us make feedback processes for writing assignments more efficient and effective.

Do I have to take part?

No, it is your choice. You do not have to participate if you do not want to. If you decide to join but then change your mind, you can stop at any time. You do not need to give us a reason if you stop.

If you decide to stop and want me to delete the data I have collected about you, please let me know by 30th June 2025.

Please know that there will be no negative consequences at all for stopping or asking your data to be deleted.

What will happen to me if I take part in the research?

If you take part in the research, you will be expected to:

- undertake a small grammar course.
- draft a brief essay on a given subject.
- submit the essay for feedback.
- use a rubric to provide feedback on a peer's essay.
- revise your essay based on the given feedback.
- undertake a survey about your experience.

The entire process is expected to take no more than ten working days, with around an hour of your time required each day. The research will be conducted within your college and no additional travel will be required.

What are the possible disadvantages and risks in taking part?

There are no risks involved in this research. Scores given in this research are not linked to your college academic career at all.

You will have to dedicate around an hour of your time each day for approximately ten working days to take part in this research. This will be during the college's working hours, possibly from 3:30 PM – 4:30 PM each day.

Are there any benefits in taking part?

This research will also give you a chance to engage in a short grammar course which may be beneficial to you. While there are no other immediate benefits for those people participating in the research, your participation can lead to insights that improve feedback processes within your college and beyond.

What information will be collected and why is the collection of this information relevant for achieving the research objectives?

We want to understand how different types of feedback affect writing. We will collect:

- Writing samples and grammar scores to assess changes
- Survey responses on your perceptions of feedback

Data will be securely stored in a University of Oxford associated OneDrive and can only be accessed by the research team – myself, my research assistant, my supervisor, and any other authorised persons within the University of Oxford.

No part of this data will be shared with anyone else in college apart from me and my research assistant. We will never reveal your personal information to anyone else.

Your college may use the collective findings of our study to make decisions about its feedback processes. However, no names or identifiable information will be given to any member of the college outside the research team.

Will the research be published? Could I be identified from any publications or other research outputs?

The findings from the research will be written up in my Master's dissertation. Your name will not be used in the dissertation. I will use direct quotations from your survey responses only if you allow.

A copy of my dissertation will be deposited online in the Oxford University Research Archive where it will be publicly available to facilitate its use in future research.

At the end of the project, group level findings from the project will be shared with your college but your individual responses will not be shared.

Data Protection

The University of Oxford is the data controller with respect to your personal data, and as such will determine how your personal data is used in the research. The University will process your personal data for the purpose of the research outlined above. Research is a task that is performed in the public interest. Further information about your rights with respect to your personal data is available from the University's Information Compliance website at https://compliance.admin.ox.ac.uk/individual-rights.

Who has reviewed this research?

This research has received ethics approval from a subcommittee of the University of Oxford Central University Research Ethics Committee. (Ethics reference: Education (Educ) DREC – 1012627**)**

Additionally, the Ethics Review Committee (ERC) at the Government Elementary College of Education (GECE), Hussainabad, Karachi has reviewed and approved this research.

Who do I contact if I have a concern about the research or I wish to complain?

If you have a concern about any aspect of this research, please contact Saira Mahmood (saira.mahmood@education.ox.ac.uk) or Dr Elizabeth Wonnacott (elizabeth.wonnacott@education.ox.ac.uk), and we will do our best to answer your query. We will acknowledge your concern within 10 working days and give you an indication of how it will be dealt with. If you remain unhappy or wish to make a formal complaint, please contact research.management@education.ox.ac.uk.

Further Information and Contact Details

If you would like to discuss the research with someone beforehand (or if you have questions afterwards), please contact:

Saira Mahmood

Master's Student,

Department of Education, University of Oxford

saira.mahmood@education.ox.ac.uk

Appendix C: CUREC Approval

Education (Educ) DREC
15 Norham Gardens, Oxford, OX2 6PY

Applicant: Saira Mahmood
Principal Investigator: Elizabeth Wonnacott
Department: Education

Study title: AI or Peer Feedback: What Works Best in Improving Writing?
(Version: 1.0)

Ethics reference: Education (Educ) DREC - 1012627

Dear Elizabeth Wonnacott,

On behalf of the Committee, I confirm that the above research study described in the application and other supporting documentation submitted to the committee has been carefully considered on behalf of the Education (Educ) DREC in accordance with the University's regulations and policy for ethics approval of research involving human participants, human tissue and/or personal data. The opinion is as follows:

Opinion of Research Ethics Committee: Favourable Opinion

Subject to the following conditions:

Decision Date: 16 Apr 2025, 19:32

Opinion End Date: 16 Oct 2026

If favourable, insurance-provided indemnity arrangements will be in place between the decision date and opinion end date and you may now commence your study activities. Should you plan to continue the research beyond the end date above, it is your responsibility to ensure that you request, and receive, an extension (via amendment) from the committee for indemnity to remain in place. You may be required to provide a justification.

Please note the following:

Amendments: Should there be any subsequent changes to the reviewed study, applications for amendments can be made via the Oxford Ethics Application System (Worktribe Ethics).

Reports: Studies considered by OxTREC are expected to submit an *annual progress report* on each anniversary of study approval, until the study is completed. An end of study report is also required.

Audit: This study may be selected for audit at the discretion of the Research Governance, Ethics and Assurance Team.

Data safety: It is the responsibility of the PI to ensure that all data collected during the course of the study is stored and transferred safely and securely in accordance with University requirements. Further guidance and advice are available from the Research Data Team. Additional information is available at https://researchsupport.web.ox.ac.uk/governance/ethics

Yours Sincerely

Education Ethics Officer

Appendix D: Writing Rubric

		Not Proficient			Proficient		
		1 Beginning	2 Emerging	3 Developing	4 Capable	5 Experienced	6 Exceptional
Content and Organization	**Thesis Statement**	Attempts to present a thesis but it is not related to the topic.	Suggests a thesis statement related to the topic, but the thesis statement is somewhat incomplete or ineffective.	Presents a complete thesis statement, but the direction of the piece is confusing	Presents an adequately clear, focused thesis statement	Develops a clear, focused, and somewhat complex and/or original thesis statement	Conveys a clear, focused, complex, and original thesis statement that drives the piece
	Details, evidence, and support	Contains evidence but it does not support the main idea	Attempts to support the main idea with some information and/ or details, but these are unfocused, unclear, and/ or unrelated	Provides incidental support of the main idea with information and/or details that lacks specificity, relevance, and/or accuracy	Supports the main idea with generally accurate, specific, and relevant information and/or details	Supports the main idea fully, with specific, credible, relevant, information and/or vivid details	Supports the main idea convincingly, with highly specific, credible information and/or striking details that go beyond the obvious

	Sequencing	Uses almost no sequencing of ideas	Uses limited sequencing that fails to show how ideas fit together	Uses sequencing that fails to showcase ideas or becomes formulaic	Provides logical sequencing of ideas	Employs sequencing that builds connections to create a unified whole	Utilizes highly effective sequencing, making best choices for progression and enrichment of reader's understanding
	Conclusion	Attempts to present a conclusion but it is not related to the topic	Conclusion is somewhat related to the topic, but it loses focuses	Conclusion is present and focused, but it fails to provide closure	Contains a conclusion that provides closure, though may be formulaic or obvious	Includes a conclusion that ties up loose ends, providing a satisfying sense of closure	Develops a satisfying conclusion that conveys a powerful and thoughtful sense of closure
Use of Language	**Transitions between paragraphs**	Almost never uses transitions between paragraphs	Rarely uses transitions between paragraphs	Uses transitions between paragraphs that are repetitive, inconsistent, and/or fail to connect ideas	Includes transitions between paragraphs that connect ideas, though they may be formulaic or predictable	Features logical, varied transitions between paragraphs that connect and develop ideas	Features thoughtful, smooth, varied transitions between paragraphs that clearly connect ideas and enhance meaning
	Vocabulary	Uses vague or limited vocabulary	Uses slightly more complex but confusing or misleading vocabulary	Uses basic, functional words that convey limited meaning	Chooses familiar vocabulary that communicates meaning	Often employs strong vocabulary that clearly conveys meaning	Consistently employs striking, powerful vocabulary that enhances meaning

							and/or shows imagination
	Sentence Structure	Uses choppy sentence structure that is incomplete, run-on, or rambling	Uses sentence structure that is simplistic and/or rarely correct	Uses sentence structure that may usually be technically correct yet isn't smooth, or overuses complicated sentences	Uses smooth, correct sentence structure that may be somewhat mechanical in places	Utilizes well-developed sentence structure throughout text	Controls strong sentence structure for maximum impact throughout text
Mechanics	**Punctuation**	Does not use or misuses punctuation nearly all the time	Uses punctuation randomly and/or incorrectly	Uses simple end punctuation that is usually correct; internal punctuation (e.g., comma, apostrophe, semicolon) contains errors	Uses mostly correct end and internal punctuation at grade level	Correctly uses both end and internal punctuation	Employs correct end and internal punctuation; may use creative punctuation that enhances readability
	Grammar	Makes serious grammar/usage , errors, making text nearly	Makes numerous grammar/usage errors, making comprehension difficult	Makes grammar/ usage errors that may distract reader; may use conversational or	Employs correct grammar/usage with few grade-level errors; minor problems	Employs proper grammar/usage	Exhibits correct grammar/ usage that contributes to clarity and style

		incomprehensible		texting language inappropriate to style	do not distort meaning or distract reader		

Adapted from Education Northwest's 6+1 Trait® Writing Model of Instruction & Assessment (Education Northwest, 2022

Appendix E: Feedback Template

FEEDBACK FORM

Your name (Reviewer):

Read **this essay and the given rubric** carefully. Your goal is to give **short, specific, and respectful** feedback to help the writer improve her work. **An example is given below.**

Example:

I think the writer did a great job with the thesis statement because she clearly outlined the two major points of the essay.

However, her writing could be even stronger if she added details to her example so that it connects better to the topic.

Now its your turn. Fill in the following form with **actionable feedback** using the given sentence structures.

1. **Content and Organization:**

I think the writer did a great job...

However, her writing could be even stronger if she...

2. **Use of Language**

I think the writer did a great job...

However, her writing could be even stronger if she...

3. **Mechanics**

I think the writer did a great job...

However, her writing could be even stronger if she...

What score would you give the essay for each of the following criteria?

Content and Organisation				**Use of Language**			**Mechanis**	
Thesis Statement	**Details, evidence and Support**	**Sequencing**	**Conclusion**	**Vocabulary**	**Transitions between paragraphs**	**Sentence Structure**	**Grammar**	**Punctuation**

Appendix F: Survey

AI vs Peer Feedback: Survey

* Required

* This form will record your name, please fill your name.

1) To what extent did revising your first draft help you improve ***your*** essay? *
   - ○ Not at all
   - ○ A little
   - ○ Somewhat
   - ○ A lot
   - ○ Extremely
2) Did you receive any feedback on your first draft? *

   Yes
   No

3) *(If Yes to Q2)* Who do you think gave you the feedback? *
   - ○ A peer (another student)
   - ○ An AI tool (like Gemini or ChatGPT)
   - ○ I'm not sure

4) What makes you think the feedback you got was from a peer or AI?

5) *(If you gave feedback to someone else)* How did giving feedback to another student help ***you*** improve ***your*** essay? *
   - ○ Not at all
   - ○ A little
   - ○ Somewhat
   - ○ A lot
   - ○ Extremely

6) Which of the following do you think would be most helpful for improving your writing? **(Select 2)** *
   - ○ Feedback from a teacher
   - ○ Feedback from peers (other students)
   - ○ Feedback from an AI tool (e.g., Grammarly, ChatGPT)
   - ○ Self-editing and practice
   - ○ Reading more examples of good writing
   - ○ One-on-one writing support
   - ○ Other

Appendix G: Selective Summary of Conceptual Feedback Typologies

Study	Typology
Ellis (2009)	Direct CF (teacher supplies the correct form) Indirect CF (error flagged, no correction) Metalinguistic CF (the teacher gives a clue about the nature of the mistake either via error codes or brief descriptions) Focus of feedback (the teacher corrects all types of linguistic errors [unfocused] or focuses on particular ones [unfocused]) Electronic feedback (teacher indicates an error and inserts a hyperlink for the student to find more information about it) Reformulation (native speaker or teacher rewrites part of the text with the student choosing to accept or not accept changes)
Cárcamo (2019)	Specification: Direct (offers correction); Localised indirect (specifies mistake; offers no correction); Unlocalised indirect (simply states number of errors) Focus: Form and function, comprehensive (one structure, many uses), unfocused (holistic) Scope: Micro (lexical-syntactic); Macro (structure, content); General (micro and macro) Source: Teacher, Classmate, Self, External (computer, researcher, native speaker, examiner) Mode of delivery (computer-mediated, handwritten) Notes: Metalinguistic (explanations or symbols); Affective (praise, emotional connection), No comments.
Pearson (2022)	Range of focus: General (overall quality); Discourse (content & organisation); Form (lexis, grammar, mechanics) Mode & tone: Advisory (suggests move); Correction (supplies fix); Criticism (negative evaluation); Description (neutral portrayal), Give-information (background data); "Need-to" (obligation cue); Praise (positive note); Question-posing (asks learner); Reader-reflection (teacher's personal reaction) Syntactic structure: declarative, imperative, exclamative, interrogative Text specificity: text-specific vs. generic Location (Marginal, Interlinear, End-comment) Explicitness: Explicit (gives revision strategy); Implicit (signals problem only); No-revision-required (evaluative/commentary)

	Length: Short (1–5 words); Average (6–15); Long (16–25); Very long (26+) Mitigation: Use of mitigation (hedging, paired-act praise-criticism, personal attribution, interrogative softeners); No mitigation Medium of delivery (pen-and-paper vs. computer-mediated) Temporality: Synchronous (while drafting); Asynchronous (after submission); Anticipatory (before issues occur)
Comajoan-Colomé & Salguero (2024)	Specification: Direct correction (supplies the correct form); Localized indirect (underline/circle to pinpoint error); Unlocalized indirect (tally of errors) Focus: Form/function (corrects one linguistic feature + a particular use); Comprehensive (corrects one linguistic feature + all uses); Unfocused (corrects a range of language forms) Scope: Micro (targets lexico-syntactic issues); Macro (targets content or structure issues); General (combines both macro and micro concerns) Source: Teacher; Classmate; Self; External (native speaker, examiner, researcher, or computer tool, etc) Mode: Computer-mediated (digital channels [e.g., LMS comments, track-changes]), Handwriting (handwritten notes) Notes: Metalinguistic explanations/symbol (codes or brief rules pointing out the nature of the error [e.g., *VT* = verb tense]), Affective (motivational or evaluative comments), None (only the corrective mark with no other comments)

Appendix H: Selective Summary of Empirically Driven Feedback Typologies

Study	Typology Created/Applied	Evidence (prevalence and/or effectiveness)
Cho et al. (2006) Creates a 6-type coding scheme to compare peer vs. expert feedback quality & perceived helpfulness	Directive (suggests a specific change to the writer's text); Nondirective (flags a problem or gives broad guidance, no fix); Praise (positive evaluation / encouragement); Criticism (negative evaluation without solution); Summary (recaps the writer's main points); Off-task (remark unrelated or ambiguous)	Expert feedback ≈ 3:1 directive; almost no summary. Undergraduates balanced directive + praise; praise ~70 % higher than expert. Writers rated directive & praise comments most helpful; nondirective neutral.
Cho and McArthur (2010) Applies a six-category feedback typology to analyse peer- versus expert comments.	Same as above	Expert (SE) feedback was 83% directive. Multiple-peer (MP) feedback had a balanced mix (Directive 46%, Praise 28%, Non-directive 11%, Summary 10%). Non-directive comments predicted more *complex-repair* and *new-content* revisions, which in turn related to quality improvement. Directive comments mainly led to simple mechanical repairs that did not raise quality.
Banihashem et al. (2024) Adopts Kerman et al., 2022 scheme to rate quality of feedback (peer vs ChatGPT) on essays	Affective (positive or negative emotion [praise, sympathy, etc.]); Cognitive–*Description* (summary of essay); Cognitive–*Identification* (pinpoint problem); Cognitive–*Justification* (explain why); Constructive (actionable recommendations)	Essays were coded on a scale: 0 (component absent), 1 (component only superficially/partially present), 2 (component fully present). Overall, components present in feedback: C-Description 1.95 > Affective 1.92 > Constructive 1.65 > C-Identification 1.41 > C-Justification 0.64. Peer >

		ChatGPT for overall. Statistically significant differences in some: ChatGPT excelled at Description ($p < .05$), while peers excelled at problem identification ($p < .01$). No category significantly linked to essay quality overall.
Lin & Crosthwaite (2024) Adapts typology from Ellis (2008) + López (2021) alongside their own to contrast teacher WCF with ChatGPT-4 WCF	Focus (Local = grammar / lexis error; Global = content & organisation issues); Scope (Focused = intensive on few issues; Unfocused = broad range); Form (Direct = supplies correct form; Indirect = does not supply correct form; Metalinguistic = provides information about the nature of errors/rules/explanation, Focused = intense feedback on a small number of issues; Unfocused = general feedback on a wide range of issues; Reformulation= rewrites sentence/text natively); Quality (Accuracy = correctness of advice; Redundancy = unnecessary or non-helpful advice)	Raw mean counts per text (teachers vs GPT) – Direct 7.18 > 3.66; Indirect 6.08 > 1.20; Reformulation 0.53 < 4.57; Metalinguistic 1.67 < 7.82. Teachers use more direct WCF (ratio ≈ 0.41 vs 0.22 across texts), but GPT's direct fixes are more *local* (0.93 vs 0.84) and slightly more *accurate* (0.987 vs 0.949). Indirect WCF is rarely used by GPT, and always local; teachers mix local & global (differences $p < .05$–$.001$). GPT exceeds teachers in reformulation and does so locally (ratio local 0.903 vs 0.498, $p < .001$) and metalinguistic feedback but also shows higher redundancy (0.057 vs 0.008, $p < .001$).