Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing?

The integration of ChatGPT as a supplementary tool for writing instruction has gained traction. However, uncertainties persist regarding how ChatGPT complements teacher assessment and the overall effectiveness of this combined approach. To address this, the study conducted a mixed-methods investigation involving 46 undergraduate students from a research university in southern China, engaging them in a Chinese academic writing task. The intraclass correlation coefficient (ICC) results revealed ChatGPT's efficiency in scoring students' writing, showing moderate to good consistency with teacher evaluations. Furthermore, a paired sample t-test unveiled significant differences in feedback quantity and types between ChatGPT and teacher assessments. Drawing from both interview data and quantitative findings, the study uncovers three ways in which ChatGPT complements teacher assessment, benefiting students with various writing proficiency levels: 1) fostering deeper comprehension of teacher assessments among students, 2) encouraging students to make judgments regarding feedback, and 3) promoting independent thinking about revisions. This study contributes to a more comprehensive understanding of the role of ChatGPT within the context of a combined assessment approach. Importantly, it underscores that certain inherent weaknesses in ChatGPT's functioning can paradoxically lead to favorable outcomes. By shedding light on the synergy between ChatGPT and teacher assessments, this research seeks to inform and enhance writing instruction in higher education.

Keywords: ChatGPT; teacher assessment; academic writing

Introduction

In recent years, a growing body of research has delved into the role of assessment in tertiary level writing instruction, recognizing it as a potent avenue for student learning (Cheong et al. 2023; Ma and Bui 2022). These assessments can be categorized based on the identity of the evaluator, including teachers, peers, self-assessment, or technology-supported assessment (Tian and Zhou 2020). Teacher assessment, the most commonly employed approach in university classrooms, entails the evaluation of students' work by teachers, involving quantitative ratings and qualitative feedback based on predefined criteria (Zheng and Xu 2023). This involves assigning numerical values to the current judgment of the writing's quality, informing students about the specific level their work has reached based on the criteria. Qualitative feedback, on the other hand, interprets these scores, offering students insights into why their work received a certain rating, identifying shortcomings, and suggesting areas for improvement. Compared to other forms of assessment, teacher assessment is often perceived as authoritative, trusted, and knowledgeable (Hu and Ren 2012; Huisman et al. 2019). It not only introduces students to effective writing strategies (Xiao and Yang 2019) but also guides them toward achieving learning objectives (Wiese and Nortvedt 2023). It has been shown to be effective in aiding students' revisions and fostering writing improvement (Tian and Zhou 2020; Cui et al. 2022).

However, teacher assessment in the university context can be labor-intensive, especially when dealing with students from various classes, imposing a substantial workload on teachers. In such scenarios, teachers may find it challenging to provide elaborated feedback and ratings to every student. This may result in hurriedly generated

feedback, such as specific solutions only, which could lead to students passively incorporating feedback into their revisions without full comprehension. This, therefore, diminishes the learning potential of the assessment (Cui et al. 2022).

This might culminate in the rapid generation of feedback, predominantly featuring specific solutions. Consequently, students might passively integrate such feedback into their revisions without a comprehensive understanding. This, in turn, diminishes the learning potential of the assessment (Cui et al., 2022).

Consequently, considerable attention has shifted to alternative or complementary assessment approaches, such as peer, self, and technology-supported assessment (Link et al. 2022; Wu et al. 2022). Among these, technology-supported assessment, particularly employing artificial intelligence (AI)-powered chatbots in writing assessment, is intriguing. These automated evaluators could potentially alleviate the workload for teachers and, in some cases, even assume part or all of the assessment responsibilities from teachers (Link et al. 2022). This shift has the potential to transform classroom assessment practices (Zhai 2022) and, consequently, reshape writing instruction and learning.

One such chatbot, ChatGPT, has garnered scholarly attention since its launch in November 2022. This chatbot engages users in conversational interactions and represents a significant leap in chatbot technology, as it is based on generative models, allowing it to provide contextual and personalized responses rather than relying on preset answers, a limitation of earlier chatbots (e.g., ELIZA-based chatbots; Guo et al. 2022; OpenAI 2023). As a result, some scholars have posited its potential for use in

writing assessment and have begun to investigate its application in this context (see Allagui 2023; Barrot 2023; Su et al. 2023). For instance, Su, Lin, and Lai (2023) provided ChatGPT with evaluation rubrics and asked it to offer feedback for text improvement. They found that ChatGPT's feedback, while promising, could sometimes be general, necessitating further clarification. Barrot (2023) tested ChatGPT's capabilities on providing a score for a given essay, suggesting that it "can" perform such work without delving into the specifics of its performance. These early yet pioneering investigations have provided initial insights into ChatGPT's utility in writing assessment.

However, the tool's inherent limitations, including the occasional provision of inaccurate, unintelligible, and/or fabricated responses (Thorp 2023), coupled with the lack of human emotional interaction (Allagui 2023), as well as limited language exposure (Mohamed 2023), may constrain its effectiveness in writing assessment. Some scholars, therefore, recommend adopting ChatGPT as a supplementary tool for writing instruction rather than relying solely on it (Barrot 2023; Mohamed 2023; Yan 2023). These insights inspired our exploration of the effectiveness of combining ChatGPT and teacher assessment in the writing classroom.

This area of study remains largely uncharted, as the integration of ChatGPT into educational settings, especially for writing, is still in its infancy. A preliminary exploration by Wei and Li (2023) provides some initial insights. They analyzed teacher and ChatGPT feedback for 208 English writing texts completed by undergraduate students from a course database, revealing a distinct focus in ChatGPT's feedback: a

tendency to offer global-level feedback, a component that was deemed beneficial for a good feedback experience but was lacking in the teacher feedback. This suggests that the deficiency identified in teacher feedback might be complemented in the feedback provided by ChatGPT, thereby enhancing the overall learning experience for students. However, their research did not involve authentic student participation, making it impossible to gauge student perceptions and actual implementations of these feedback forms during revisions. This gap leaves us without a clear understanding of how these two feedback sources might synergize effectively and efficiently, ultimately contributing to writing improvement. Hence, further research into the fusion of ChatGPT and teacher assessment, with a specific focus on ChatGPT's role in this amalgamation, is indispensable. This research could offer teachers valuable insights into the seamless integration of AI-based tools into traditional teaching assessment practices, ultimately optimizing the instructional process.

Informed by the literature, this study employs the terms "teacher assessment" and "ChatGPT assessment" to refer to the processes of assigning scores and providing feedback based on identical assessment criteria. This investigation into the effectiveness of combining these two assessments considers three primary facets: 1) the features of the two assessments, 2) students' perceptions of the feedback embedded in these assessments, and 3) students' actual implementation of the feedback embedded in these assessments. The first facet delves into the unique characteristics of ChatGPT and teacher assessments and elucidates how the uniqueness of each assessment can be optimally leveraged, while the latter two aspects place a focal lens on the feedback,

endeavoring to ascertain whether students utilize ChatGPT feedback and to what extent ChatGPT, in conjunction with teacher feedback, assists students in enhancing their writing. This study is guided by the following research questions:

RQ1: What are the features of scores and feedback provided by teachers and ChatGPT in the assessment of students' writing?

RQ2: How do students perceive teacher and ChatGPT feedback?

RQ3: How do students implement teacher and ChatGPT feedback to facilitate their revisions?

Methodology

Participants

A total of 46 Chinese undergraduate students, comprising 4 males and 42 females, were involved in this study (M_{age} = 20.35, SD = 1.48). They were all education-focused majors and were recruited from a research university in southern China through convenience sampling. These students were enrolled in a six-week academic Chinese writing training program, with three-hour weekly sessions. The program aimed to provide comprehensive instruction in various forms of academic writing and to enhance the writing skills of the participants. Participants provided informed consent and received 35 RMB (Renminbi, the Chinese currency) compensation for their full program participation. They were designated as S1 to S46 in the subsequent sections.

Materials

Writing task and assessment criteria

In this study, students were tasked with composing a Chinese abstract of approximately 300 words for a specified empirical article. This form of academic writing was chosen because it requires students to organize, select, and integrate information from the source article, thereby enhancing competence in discourse synthesis, a fundamental skill in academic writing (Nelson and King 2023). The chosen article, composed in Chinese and lacking an abstract, centered on the contemporary challenges in Chinese language teaching classrooms. Given that the content was familiar to all participants majoring in education-focused fields, and no complex statistical knowledge was necessary for comprehension, the authors collectively deemed the material appropriate for students enrolled in this writing training program.

Abstracts were evaluated on five dimensions: (a) research purpose, (b) method, (c) findings, (d) implications, and (e) language conventions (Lu et al. 2021). The criteria followed the common "IMRaD" structure for social sciences academic writing (Tabuena 2020) and emphasized readability (Gazni 2011). Since both the ability to organize the content logically (reflected in the first four dimensions) and express the content concisely (the last dimension) are important for academic writing, we assigned equal weight to each dimension. Each dimension had five levels to assess strengths and weaknesses, resulting in scores from zero to eight, with a maximum total score of 40. The criteria were refined based on suggestions from two writing experts, the sixth author and the program teacher.

Assessment form

An assessment form was devised to offer students both teacher and ChatGPT assessments. This form comprises three columns, each with a specific purpose. The leftmost column outlines the criteria for the writing task, providing students a clear framework to review the assessments they received. The second column pertains to teacher assessment and encompasses three subsections: (1) a score for each dimension, (2) a justification for each score (i.e., what do you think of the students' abstract drafts?), and (3) suggestions for enhancing performance on each dimension (i.e., how the students can revise their compositions for enhanced performance?). The third column is exclusively for ChatGPT, where it emulates the teacher's role by scoring, justifying scores, and suggesting improvements in alignment with the specified dimensions (See Appendix A). To mitigate potential biases stemming from student preferences and/or column sequencing, both the second and third columns are consistently denoted as Assessment 1 and Assessment 2, concealing the origins of each assessment. Furthermore, half of the forms assign teacher assessments as Assessment 1 and ChatGPT's as Assessment 2, while the other half reverses this order.

Preparation for ChatGPT assessment

ChatGPT version 3.5 was used for its easy accessibility. It is worth highlighting that ChatGPT is capable of generating responses based on prompts (Kasneci et al. 2023). To make ChatGPT work like the program teacher in this study, OpenAI's guidelines (2022) for ChatGPT usage were closely followed. We meticulously crafted the

following prompts in Chinese to ensure consistency with those furnished to the teachers in the assessment form. These three prompts underwent pilot testing using two abstracts written by research assistants, with subsequent minor adjustments:

1. “In your role as a professional Chinese academic writing teacher, you administered a writing task for students. This task required students to read a specific article and then compose an abstract of approximately 300 words. Your responsibility is to assess the students’ abstract writing. Before commencing the actual assessment, you are required to read the article used for the task. Please provide a response indicating your understanding of the article after you have completed your reading. <Insert the article here>.”
2. “Recall the article you have just read. Now, review the assessment criteria employed for this task. Provide a response to indicate your comprehension of the criteria following your review. <Insert the criteria here>.”
3. “Drawing from the article and criteria you have reviewed, you are now prepared to assess the students’ abstract writing. Kindly complete the following three steps in order: (1) assign a score to evaluate the students’ writing performance on each dimension as per the criteria, (2) provide a justification for each score to elucidate what you think of the students’ abstract drafts, and (3) offer suggestions for improvement on each dimension to indicate how the students can revise their compositions for enhanced performance. <Insert the student’s abstract here>.”

It is pertinent to note that prompts one and two were employed once, while prompt three was recurrently used to assess different students’ writing. However, after the initial use of prompt three for the first student, a minor adjustment was made to the opening phrase: “Ok, now you continue to assess the next student's abstract writing. Kindly complete...”.

Procedure

This study comprised two sessions over two weeks. In week 1, the program teacher conducted a 50-minute session teaching students how to write an academic abstract and use assessment criteria. This included presenting an exemplar abstract and facilitating a discussion and rating activity among students to enhance their comprehension of abstract writing and criteria.

In the subsequent phase, students were allocated 80 minutes to peruse a designated article, compose an abstract, and subsequently submit their work via the online platform. In the following week, students were administered with an assessment form, encompassing teacher and ChatGPT assessments. Within a 60-minute timeframe, they were tasked with refining their initial drafts based on the provided assessments. The revised abstracts were subsequently resubmitted through the online portal.

Data collection and analysis

The data collection involved acquiring textual data, encompassing the initial and revised abstracts, as well as assessment forms that included assessments from teachers and ChatGPT. In addressing RQ1, particular focus was on the scores provided by teachers and ChatGPT, alongside the distinct features of their feedback. Expert evaluators, the fifth author and a program teacher experienced in academic writing instruction, independently assessed the students' initial and revised abstracts based on predefined criteria. A meeting was held before scoring to ensure shared criteria

understanding. Both evaluators then independently assessed all the initial and revised abstracts, showing high reliability (Intraclass correlation coefficient, ICC = 0.90 for both initial and revised abstracts; Cicchetti 1994). Writing improvement was measured by the score difference between revised and initial abstracts.

In coding teacher and ChatGPT feedback on assessment forms, two coders, the first author and a research assistant with a master's degree, engaged in comprehensive discussions and conducted trial coding with three randomly selected sets of feedback from both sources. Subsequently, they independently coded the remaining sets. Any discrepancies were resolved through discussion. Inter-rater reliability was determined using Cohen' s kappa. Appendix B provides a detailed coding process for teacher and ChatGPT feedback. Two feedback features, feedback quantity and types, were examined, in line with recent research (Lu et al. 2021; Wu and Schunn 2023). To determine the feedback quantity, feedback comment in each dimension was first segmented into independent units. A unit is independent information addressing a singular issue in the text (Wu and Schunn 2023). Feedback quantity was calculated by summing the units across five dimensions. Feedback types included (a) summary, (b) praise, (c) explanation, (d) specific solution, and (e) general suggestion. Their presence was coded within each dimension on the assessment form, with a presence assigned a value of 1. The quantity of each feedback type was determined by summing the presence of each type across the five dimensions. Paired sample t-tests were used to analyze differences in feedback quantity and types between teacher and ChatGPT feedback.

To examine RQ2, interviews were conducted to explore students' perceptions of teacher and ChatGPT feedback. Representative students were selected based on their average scores for initial and revised abstracts (M = 24.3, SD = 4.74), leading to the formation of high (M = 29.4; N = 14), medium (M = 24.8; N = 16), and low (M = 19.3; N = 16) writing level groups (Cheong et al. 2023). Three students from each group were randomly chosen for interviews conducted in Chinese. During the interviews, students were prompted to: (1) review teacher and ChatGPT feedback, (2) share their perceptions regarding understanding, agreement, and preference for the two feedback, and (3) describe their revision processes for initial abstracts. Questions (e.g., how much did you understand teacher feedback?) adapted from relevant research were used to encourage detailed responses (Zhu et al. 2023). Each participant received an additional 5 RMB as compensation, and interviews were scheduled a week after students submitted their revised abstracts. The interviews, lasting around 30 minutes each, were audio-recorded with student consent and transcribed verbatim for analysis by the sixth author and a research assistant.

To investigate RQ3, we focused on students' implementation of teacher and ChatGPT feedback for their revisions, their revision behaviors, and the resulting writing improvement. The coders responsible for feedback coding also undertook revision coding, commencing with a discussion and a trial coding session before the actual coding. Four steps were taken to determine the actual feedback implementation. First, we identified the quantity of implementable feedback in each dimension on the assessment form. Feedback categorized as a problem, specific solution, and/or general

suggestion was considered implementable (Wu and Schunn 2021). Second, we compared the initial and revised abstracts using Micro Soft Word's "Compare Documents" function to trace changes, excluding format alterations. Third, each change source was scrutinized by pinpointing relevant implementable comments in either teacher or ChatGPT feedback. Fourth, we coded these implementable comments as fully implemented, partially implemented, or rejected in each dimension. The implemented rate of feedback in each dimension was calculated as the percentage of the sum of fully and partially implemented comments within the implementable feedback.

To elucidate students' revision behaviors, each revision in each dimension was categorized into one of five patterns: revision based on teacher feedback only, ChatGPT feedback only, based on both, extra revision, or no revision. Notably, the "based on both" and "extra revision" patterns may coexist, indicating that students utilized both feedback and initiated revisions beyond feedback themselves. To capture students' writing improvement triggered by the revisions, we employed a split-three procedure, classifying improvement into three scales: major (increase ≧ 9.5 marks; N = 16), moderate (5 marks < increase < 9.5 marks; N =14), and minor (0 < increase ≦ 5 marks; N = 13) (Cheong et al. 2023). The improvement was matched with each revision behavior. Appendix C provides details and examples of feedback implementation and revision coding. We primarily utilized SPSS version 26.0 for conducting descriptive statistics, computing Intraclass Correlation Coefficient (ICC), assessing Cohen's kappa, and performing paired sample t-tests.

Results

Features of teacher and ChatGPT assessments

To address RQ1, we computed Intraclass Correlation Coefficients (ICCs) to assess the consistency between the scores provided by teachers and ChatGPT. Table 1 illustrates the results. Most ICC values ranged from 0.6 to 0.75, with a few exceeding 0.75, indicating moderate to good consistency between teacher and ChatGPT scores (Koo and Li 2016). This suggests that ChatGPT can effectively score students' writing, similar to teachers.

<insert Table 1>

Furthermore, paired sample *t*-tests were conducted to explore differences in feedback features between teachers and ChatGPT. Table 2 presents the results, indicating significant differences in both feedback quantity and various types. ChatGPT feedback displayed a higher feedback quantity (t (45) = 3.73, p = 0.001, Cohen's d = 0.55), as well as more summary (t (45) = 3.24, p = 0.002, Cohen's d = 0.48) and general suggestions (t (45) = 8.73, p < 0.001, Cohen's d = 1.29) compared to teacher feedback. Conversely, teacher feedback featured a greater quantity of praise (t (45) = 7.44, p < 0.001, Cohen's d = 1.10), explanation (t (45) = 4.32, p < 0.001, Cohen's d = 0.64), and specific solution (t (45) = 5.86, p < 0.001, Cohen's d = 0.86) types. The Cohen's d values, ranging from 0.48 to 1.29, signify a medium to large practical significance of the results (Cohen 1988). These results emphasize that ChatGPT is capable of generating feedback distinct from that provided by teachers.

<insert Table 2>

Students' perceptions of teacher and ChatGPT feedback

To address RQ2, students' perceptions of teacher and ChatGPT feedback were primarily reflected in the interview data. All nine students expressed greater understanding and agreement with teacher feedback compared to ChatGPT feedback:

> "I find Feedback 1 (teacher feedback)[1] easier to understand since it not only clearly identifies my weaknesses but also provides specific advice. In contrast, Feedback 2 (ChatGPT feedback) consists of lengthy comments that I find difficult to comprehend and apply." (S3)

Seven students mentioned readability, clarity, and specificity as the primary factors that made teacher feedback more comprehensible. Consequently, they were more prone to agree with teacher feedback that they could readily grasp. Furthermore, whether the feedback aligned with their existing thoughts played a role in their agreement, as articulated by S5:

> "I tend to agree with Feedback 1 (teacher feedback) as it addresses my concerns. However, I suspect Feedback 2 (ChatGPT feedback) didn't thoroughly read my text. It suggested adding information about instructional strategies, which I believe I already included in my initial draft. Upon careful reading the provided article and my text, I question the accuracy of Feedback 2." (S5)

Regarding feedback preferences, students from different writing level groups offered varying perspectives. Two interviewees from low-level groups (S3, S5, and

[1] Students were not informed of the origins of the feedback due to the structure of the assessment form. Nevertheless, for the sake of clarity, we have presented the sources of feedback here.

S37), who demonstrated lower understanding and agreement with ChatGPT feedback, also expressed a lower preference for this type of feedback. They found ChatGPT feedback to contain esoteric statements, vague explanations of text issues, and general or inaccurate revision advice, which made them less receptive to it. As S37 explained:

> "I feel that Feedback 1 (ChatGPT feedback) wasn't suitable for me. Given my relatively low writing level, I might struggle to improve if the feedback doesn't clearly instruct me on how to revise my writing." (S37)

In contrast, students from high-level writing groups (S15, S27, and S39) showed a preference for ChatGPT feedback. They believed that this type of feedback served as a source of inspiration as it did not directly provide explicit revision directives. Instead, it motivated them to independently reflect on their writing and autonomously elaborate on revisions:

> "I found Feedback 2 (ChatGPT feedback) quite inspiring. It guided me towards potential revisions instead of providing explicit instructions. I was intrigued to explore how to implement the guidance. Consequently, I carefully reflected my text, the criteria, and generated new ideas that diverged from both feedback to enhance the quality of my writing." (S15)

Another student from the same group expressed a preference for ChatGPT feedback for its contribution to developing the capacity to make judgments about feedback. This is closely relevant to feedback literacy, an essential skill for optimizing learning from feedback (Yu and Liu 2021; Zhang et al. 2023).

> "Feedback 1 (ChatGPT feedback) suggested adding more description about research backgrounds, while Feedback 2 (teacher feedback) advised reducing the description about research backgrounds. It is intriguing to encounter these conflicting suggestions as it motivates me to evaluate which feedback might benefit my writing improvement. After conducting a careful pairwise comparison between my text, the established criteria, and the two sets of feedback, I decided to adopt Feedback 1 and added statements beyond what it suggested." (S39)

It is noteworthy that both kinds of feedback can jointly encourage students to make judgments about feedback. Indeed, students from medium-level writing groups (S10, S21, and S43) indicated that both feedback were necessary and complemented each other in promoting understanding:

> "I believe that the two kinds of feedback work best in concert. Although Feedback 1 (teacher feedback) told me on revising the concluding sentence, it was only when I reviewed the writing principles and criteria outlined in Feedback 2 (ChatGPT feedback) that I understood why I needed to make this change and recognized the usefulness of Feedback 1's advice. Moreover, Feedback 2 reminded me to consult the criteria when making revisions." (S21)

In summary, these findings reveal a disparity between students' understanding and agreement with feedback and their preferences. Students exhibit complex preferences for the two kinds of feedback, with differing writing levels yielding different perspectives regarding the role of ChatGPT feedback. Preference often hinges on how students perceive the benefits of the feedback. Consequently, in some cases, even when feedback provides implicit instructions that may result in lower understanding, it can evoke a high preference when students recognize its potential

benefits, as exemplified by S15's experience. Additionally, these results suggest that teacher and ChatGPT feedback can complement each other, a topic we would explore further in the following discussion section.

Students' implementation of teacher and ChatGPT feedback for revisions and writing improvement

To address RQ3, we examined students' implementation of both teacher and ChatGPT feedback in their revisions, their revision behaviors, and the ensuing improvements. Teacher feedback was implemented more frequently than ChatGPT feedback in all dimensions, with a total implementation rate of 67.6% compared to ChatGPT's 39.6% (see Table 3). ChatGPT feedback had a higher rejection rate at 40.1%, while teacher feedback was rejected 19.8% of the time. In general, students tended to implement teacher feedback more than ChatGPT's during revisions.

<insert Table 3>

In terms of revision behavior, 170 out of 261 cases (65.1%) involved students using both teacher and ChatGPT feedback for their revisions (see Table 4). In the remaining cases, extra revisions, often mentioned in students' interviews alongside ChatGPT feedback, were the most common (19.2%), followed by revisions solely based on teacher feedback (13.4%). This indicates that the primary revision approach of students was to integrate both teacher and ChatGPT feedback.

<insert Table 4>

As previously introduced, students' improvements were categorized into three scales to gauge how each improvement scale corresponded with specific revision behaviors. The results indicated that 215 out of 234 cases (91.9%) saw students' scores improve (see Table 5). Specifically, their scores improved saliently in 87 instances following all types of revisions, moderately in 81 cases, and slightly in 47 cases. Notably, combining both teacher and ChatGPT feedback was the most instances in all improvement categories. Extra revisions and teacher feedback-only revisions followed. This emphasizes the crucial role of integrating both feedback sources in enhancing writing performance.

<insert Table 5>

Discussion

In a mixed-methods investigation, this study delved into the combined use of teacher and ChatGPT assessments in the context of Chinese academic writing. The major findings are discussed below.

Discrepancies in features between teacher and ChatGPT assessments

In response to RQ1, the analysis of ICCs revealed that ChatGPT can effectively score students' writing, displaying moderate to good consistency with teacher provided scores. This not only supports Barrot's assertion in 2023 that ChatGPT "can" perform assessment tasks but also emphasizes that its performance is acceptable. Additionally, the subsequent results of paired sample t-tests indicated distinctions between ChatGPT

and teacher feedback in terms of both quantity and types. The finding that more general suggestions were included in the ChatGPT feedback types compared to teacher feedback types is noteworthy. It echoes the observations of Allagui (2023) and Su, Lin, and Lai. (2023), both of whom found ChatGPT feedback to be somewhat abstract and general, lacking clear and specific recommendations for improvement. This could be attributed to ChatGPT's inherent limitations in logical reasoning, as it tends to "guess" the meaning of user input rather than seeking clarification. Consequently, it may produce responses that are nonsensical, fake, or overly abstract (Su et al. 2023; Thorp 2023).

Another discernible deficiency in ChatGPT's capabilities is its absence of human-like emotional interaction. Our findings reveal that ChatGPT feedback contained significantly less praise in comparison to teacher feedback, which partly concurs with Allagui's (2023) observation that ChatGPT rarely offers responses praising for students' interesting writing. Feedback related to emotional aspects is pivotal in nurturing social interactions between feedback providers and students. Unlike feedback on cognitive aspects (e.g., specific solutions), positive emotional feedback holds the potential to inspire a positive perception in students, motivating them to improve their writing performance (Lu et al. 2021, 2023). From this perspective, ChatGPT's limitation in establishing emotional connections with students through text highlights the view of technical rationality in education, which underscores the value of technology-enhanced teaching while acknowledging that

technology cannot wholly replace human teaching (Farazouli et al. 2023; Mohamed 2023).

The role of ChatGPT assessment in conjunction with teacher assessment

In response to RQ2 and RQ3, we sought to investigate how students perceive and implement the feedback in these two assessments. The results provide empirical support for the combined use of teacher and ChatGPT feedback in the Chinese academic writing. The substantial implementation rates of both teacher feedback (80.2%) followed by ChatGPT feedback (59.9%) underscore the central role of teacher assessment and the complementary role of ChatGPT assessment (Table 3). This aligns with the argument that ChatGPT serves best as a supplementary tool in writing instruction (Barrot 2023).

Specifically, ChatGPT assessment can complement teacher assessment in several ways. First, ChatGPT assessment aids in students' understanding of teacher assessment and vice versa. RQ1 results highlighted differences in the types of feedback provided. While each type has its unique role, together, they offer students coherent information about their current writing performance, weaknesses, possible revision strategies, and motivation. The absence of either type can lead to insufficient information for students to understand the feedback's purpose, potentially impacting feedback implementation (Patchan et al. 2016; Lu et al. 2021). Thus, the amalgamation of these two kinds of feedback is more likely to offer a comprehensive set of information, addressing potential information gaps present in any single source. This,

in turn, enhances students' mutual understanding of these feedback sources. For example, an interview with S21 revealed that explanations in ChatGPT feedback helped the student understand the specific solutions in teacher feedback. Except in cases where the feedback types in different feedback sources could be complementary and enhance feedback understanding, when the two kinds of feedback coincide on a particular issue, they can reinforce the credibility of that issue, further promoting mutual understanding.

Second, ChatGPT assessment encourages students to exercise judgment about their feedback. Despite the inherent limitations of ChatGPT, such as providing incorrect, irrelevant, or abstract responses (Thorp 2023), it paradoxically stimulates positive washback. The study identified two typical situations where ChatGPT feedback triggered students to make judgments, a critical aspect of feedback literacy for optimizing learning (Zhang et al. 2023). In the first situation, when ChatGPT feedback conflicted with teacher feedback, students had to evaluate the applicability of both feedback kinds, deciding the appropriate usage of them, as reported by S39. In the second situation, when ChatGPT feedback was incorrect, students had to detect and diagnose these errors, as reported by S5. In both situations, students may engage in frequent cross-comparisons between the original article, the abstract text, the two types of feedback, and assessment criteria. This process enhanced their understanding of their current writing, assessment criteria, as well as the value and role of different feedback, thereby improving their capacity to make informed judgments about feedback.

Third, ChatGPT assessment complements teacher assessment by encouraging students to think independently about revisions. Encouragingly, approximately a quarter of revisions (19.2%) constituted “extra revisions” (Table 4). This kind of revision positively contributed to writing improvement (Table 5), consistent with the findings of previous research by Lu, Yao, and Zhu (2023), which associated extra revision with enhanced writing performance. ChatGPT, primarily offering general suggestions in its feedback due to inherent limitations, may play a pivotal role in stimulating these additional revisions. These suggestions are not typically problem-oriented but encompass statements related to possible revision directions, writing principles, approaches, and/or criteria. In contrast to detailed solutions, which can lead to passivity and dependency on provided answers (Strijbos et al. 2010), general suggestions encourage students to elaborate on the information provided, promoting deeper reflection on their revisions. This process ultimately leads to the discovery of their own revision solutions. S15, for instance, reported such an experience in the interview, where she meticulously reviewed the text and criteria to elaborate on the revision direction provided by ChatGPT in her unique way. Moreover, the two typical situations—conflicts between ChatGPT and teacher feedback and detecting errors in ChatGPT feedback—can also spark such revisions. The frequent comparisons students engaged in between the text and the two feedback sources enabled them to identify gaps between their desired and current writing, and inconsistencies between their text and criteria, fostering independent thinking during revisions.

In addition, it is noteworthy that students of all writing levels can benefit from the combination of ChatGPT assessment with teacher assessment. Table 5 indicated that revisions based on both ChatGPT and teacher feedback were the most prevalent across all writing improvement categories, demonstrating the effectiveness of this combined approach. Notably, different levels of students benefit in various ways. Interviews revealed that high-level students often benefited from reflections on their revisions, largely inspired by potential weaknesses identified in ChatGPT feedback. Medium-level students found value in using ChatGPT feedback to comprehend teacher feedback, while lower-level students primarily benefited from teacher feedback, only resorting to ChatGPT feedback if it provided clear, easily understandable revision guidance. Research on the extent to which different student writing levels can benefit from this approach is crucial, as it relates to how students can engage with this unique writing assessment paradigm.

In summary, ChatGPT assessment, when combined with teacher assessment, plays a valuable role in supporting students' understanding of feedback, judgment of feedback, independent thinking about revisions, and offers benefits across diverse student writing levels. This approach thus optimizes the use of AI-based tools in traditional assessment practices.

Conclusion

By exploring the features of teacher and ChatGPT assessment, as well as students' perceptions and implementations of the feedback in these two assessments, this study

contributes to the existing literature in two key aspects. Firstly, it extends the understanding of teacher and ChatGPT assessment within the context of writing by highlighting their distinct features and showcasing the effectiveness of their combination in enhancing students' writing revisions and subsequent improvements. Secondly, it emphasizes the role of ChatGPT within this combined approach, particularly highlighting that some of its inherent weaknesses can produce positive outcomes, such as stimulating students' reflection on their writing. This highlights the perspective that the utility of technology depends on how it's employed rather than its inherent weaknesses, aligning with the view presented by Yao et al. (2021), as well as Link et al. (2022).

It is recommended that teachers consider incorporating ChatGPT as a supplementary tool for writing assessment. They can utilize ChatGPT's scoring and feedback as a reference for their own assessments, thereby reducing the workload and leveraging the strengths of both approaches. Then, teachers should provide students with specific guidance on how to effectively utilize ChatGPT and teacher assessment. This guidance can include teaching students the skills needed to conduct comparisons between their text, assessment criteria, and the different feedback sources to evaluate the applicability of diverse feedback and identify potential revision strategies. Moreover, tailored assistance should be offered to students at various writing proficiency levels. Low-level students may benefit from instruction on how to respond when ChatGPT provides general, incorrect, or nonsensical feedback. In contrast, medium and high-level students should be encouraged to engage in reflective revision,

exploring opportunities for extra revisions. This approach can help cultivate their independence and autonomy as learners and thinkers.

However, several limitations of this study should be noted. Firstly, it relied on a small sample of student writings within a specific genre. Future research should encompass a more extensive sample, spanning various genres such as proposals, to further investigate the application of AI in academic writing for tertiary level students. Additionally, the response of ChatGPT is heavily influenced by the prompt input. Some of the deficiencies found in this study, like the lack of praise in ChatGPT's feedback, may be linked to the absence of explicit instructions in the prompts. It is possible that with clearer directives, ChatGPT could provide the desired elements. Therefore, future research might consider experimenting with different prompts to gain insights into how to effectively train this tool for writing assessment applications.

Data availability statement

The data are available from the corresponding author upon reasonable request.

Ethical statement

The studies involving human participants were reviewed and approved by the university where the study conducted. The participants provided their written informed consent to participate in this study.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Allagui, B. 2023. "Chatbot Feedback on Students' Writing: Typology of Comments

and Effectiveness." In *International Conference on Computational Science and Its Applications*, 377-384. Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-37129-5_31

Barrot, J. S. 2023. "Using ChatGPT for second language writing: Pitfalls and potentials." *Assessing Writing*, *57*, 100745. doi: 10.1016/j.asw.2023.100745

Cicchetti, D. V. 1994. "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology." *Psychological assessment*, *6*(4): 284. doi: 10.1037/1040-3590.6.4.284

Cohen, J. 1988. "*Statistical Power Analysis for the Behavioral Sciences*." Hillsdale, NJ: Erlbaum.

Cui, Y., Schunn, C. D., and Gai, X. 2022. "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. doi: 10.1080/07294360.2021.1969541

Farazouli, A., Cerratto-Pargman, T., Bolander-Laksov, K., and McGrath, C. 2023. "Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices." *Assessment & Evaluation in Higher Education*, 1-13. doi: 10.1080/02602938.2023.2241676

Gazni, A. 2011. "Are the abstracts of high impact articles more readable? Investigating the evidence from top research institutions in the world." *Journal of Information Science*, *37*(3): 273-281. doi: 10.1177/0165551511401658

Guo, K., Wang, J., and Chu, S. K. W. 2022. "Using chatbots to scaffold EFL students' argumentative writing." *Assessing Writing*, *54*, 100666. doi: 10.1016/j.asw.2022.100666

Hu, G. and Ren, H. 2012. "The impact of experience and beliefs on Chinese EFL student writers' feedback preferences." In *Academic Writing in a Second or Foreign Language*, edited by Tang R , 67–87. London, UK: Continuum.

Huisman, B., Saab, N., van den Broek, P., and van Driel, J. 2019. "The impact of formative peer feedback on higher education students' academic writing: a Meta-Analysis." *Assessment & Evaluation in Higher Education*, *44*(6): 863-880.

doi: 10.1080/02602938.2018.1545896

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... and Kasneci, G. 2023. “ChatGPT for good? On opportunities and challenges of large language models for education.” *Learning and individual differences*, *103*, 102274. doi: 10.1016/j.lindif.2023.102274

Koo, T. K., and Li, M. Y. 2016. “A guideline of selecting and reporting intraclass correlation coefficients for reliability research.” *Journal of chiropractic medicine*, *15*(2): 155-163. doi: 10.1016/j.jcm.2016.02.012

Link, S., Mehrzad, M., and Rahimi, M. 2022. “Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement.” *Computer Assisted Language Learning*, *35*(4): 605-634. doi: 10.1080/09588221.2020.1743323

Lu, Q., Yao, Y., and 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. doi: 10.1016/j.asw.2023.100725

Lu, Q., Zhu, X., and Cheong, C. M. 2021. “Understanding the difference between self-feedback and peer feedback: A comparative study of their effects on undergraduate students' writing improvement.” *Frontiers in psychology*, *12*, 739962. doi: 10.3389/fpsyg.2021.739962

Ma, M., and Bui, G. 2022. “Implementing continuous assessment in an academic English writing course: An exploratory study.” *Assessing Writing*, *53*, 100629. doi: 10.1016/j.asw.2022.100629

Mohamed, A. M. 2023. “Exploring the potential of an AI-based Chatbot (ChatGPT) in enhancing English as a Foreign Language (EFL) teaching: perceptions of EFL Faculty Members.” *Education and Information Technologies*, 1-23. doi: 10.1007/s10639-023-11917-z

Nelson, N., and King, J. R. 2023. “Discourse synthesis: Textual transformations in writing from sources.” *Reading and Writing*, *36*(4): 769-808. doi: 10.1007/s11145-021-10243-5

OpenAI. 2022, November 30. “Introducing ChatGPT.” https://openai.com/blog/chatgpt

OpenAI 2023, March 1. “Introducing ChatGPT and Whispers APIs.” https://openai.com/blog/introducing-chatgpt-and-whisper-apis

Patchan, M. M., Schunn, C. D., and Correnti, R. J. 2016. “The nature of feedback: How peer feedback features affect students’ implementation rate and quality of revisions.” *Journal of Educational Psychology*, *108*(8): 1-23. doi: 10.1037/edu0000103

Strijbos, J. W., Narciss, S., and Dünnebier, K. 2010. “Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency?.” *Learning and instruction*, *20*(4): 291-303. doi: 10.1016/j.learninstruc.2009.08.008

Su, Y., Lin, Y., and Lai, C. 2023. “Collaborating with ChatGPT in argumentative writing classrooms.” *Assessing Writing*, *57*, 100752. doi: 10.1016/j.asw.2023.100752

Tabuena, A. C. 2020. “Students’ perception in the implementation of the IMRaD structure approach and its implications on the research writing process.” *International Journal of Research Studies in Education*, *9*(7): 55-65. doi: 10.5861/ijrse.2020.5913

Tian, L., and Zhou, Y. 2020. “Learner engagement with automated feedback, peer feedback and teacher feedback in an online EFL writing context.” *System*, *91*, 102247. doi: 10.1016/j.system.2020.102247

Thorp, H. H. 2023. “ChatGPT is fun, but not an author.” *Science*, *379*(6630): 313-313. doi: 10.1126/science.adg7879

Wei, S., and Li, L. Y., 2023. “Artificial intelligence-Assisted Second Language Writing Feedback: A Case Study of ChatGPT.” *Foreign Languages in China*, *20*(3): 33-40. doi:10.13564/j.cnki.issn.1672-9382.2023.03.007

Wiese, E., and Nortvedt, G. A. 2023. “Teacher assessment literacy in culturally and linguistically diverse classrooms: A Norwegian case study.” *Teaching and Teacher Education*, *135*, 104357. doi: 10.1016/j.tate.2023.104357

Wu, W., Huang, J., Han, C., and Zhang, J. 2022. “Evaluating peer feedback as a reliable and valid complementary aid to teacher feedback in EFL writing classrooms: A feedback giver perspective.” *Studies in Educational Evaluation*, *73*, 101140. doi: 10.1016/j.stueduc.2022.101140

Wu, Y., and 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. doi: 10.3102/0002831220945266

Wu, Y., and Schunn, C. D. 2023. “Passive, active, and constructive engagement with peer feedback: A revised model of learning from peer feedback.” *Contemporary Educational Psychology*, *73,* 102160. doi: 10.1016/j.cedpsych.2023.102160

Xiao, Y., and Yang, M. 2019. “Formative assessment and self-regulated learning: How formative assessment supports students' self-regulation in English language learning.” *System*, *81*, 39-49. doi: 10.1016/j.system.2019.01.004

Yan, D. 2023. “Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation.” *Education and Information Technologies*, 1-25. doi: 10.1007/s10639-023-11742-4

Yao, Y., Wang, W., and Yang, X. 2021. “Perceptions of the inclusion of automatic writing evaluation in peer assessment on EFL writers’ language mindsets and motivation: A short-term longitudinal study.” *Assessing Writing*, *50*, 100568. doi: 10.1016/j.asw.2021.100568

Yu, S., and Liu, C. 2021. “Improving student feedback literacy in academic writing: An evidence-based framework.” *Assessing Writing*, *48*, 100525. doi: 10.1016/j.asw.2021.100525

Zhai, X. 2022. “ChatGPT User Experience: Implications for Education.” SSRN Scholarly Paper. Rochester, NY. doi:10.2139/ssrn.4312418. doi.org/10.2139/ssrn.4312418

Zhang, E. D, Liu, C., and Yu, S. 2023. “The impact of a feedback intervention on university students’ second language writing feedback literacy.” *Innovations in Education and Teaching International,* 1-17. doi:

10.1080/14703297.2023.2254275

Zheng, Y., and Xu, J. 2023. “Unpacking the impact of teacher assessment approaches on student writing engagement: a survey of university learners across different languages.” *Assessment & Evaluation in Higher Education*, 1-14. doi: 10.1080/02602938.2023.2219431

Zhu, W., Yu, S., and Zheng, Y. 2023. “Exploring Chinese EFL undergraduates’ academic emotions in giving and receiving peer feedback on writing.” *Assessment & Evaluation in Higher Education*, 1-17. doi: 10.1080/02602938.2023.2235635