Research Article

TEFL Ph.D. Students' Engagement with Peer Feedback and Automated Feedback on their Dissertation Proposals

Lotfollah Akbarpour[1], Leila Saberi[2*], Saeed Yazdani[3]

[1,3]*English Department, Bushehr Branch, Islamic Azad University, Bushehr, Iran*
[2]*Department of English, Marvdasht Branch, Islamic Azad University, Marvdasht, Iran*

*Corresponding author: saberi500@yahoo.com

(Received: 2023/12/11; Accepted: 2025/01/28)
Online publication: 2025/01/31

Abstract

This study was conducted to examine the comments provided by ChatGPT 3.5 and doctoral students on doctoral dissertation proposals. Feedback receivers' behavioral engagement with these two feedback types was also examined. The participants of this study, selected based on convenience sampling, were 28 Teaching English as a Foreign Language PhD students from three provinces who wrote their dissertation proposals in English. The first version and revised versions of their proposals and the provided comments were analyzed to identify the feedback types and the extent they applied the comments. Furthermore, stimulated recall interviews were used to identify the reasons why they did not apply some of the participants. The findings showed that both ChatGPT 3.5 and doctoral students were successful in providing both content-related and form-related comments. These two feedback sources also provided significant numbers of elaborated and justified feedback on dissertation proposals. The feedback receivers applied most of the comments provided by these feedback sources, and the specificity levels of comments affected the incorporation rate. The findings of the thematic analysis of the stimulated recall interview data revealed that the participants did not apply the comments because they were too broad, inaccurate, difficult to apply, and difficult to understand.

Keywords: ChatGPT 3.5-generated feedback, peer feedback, academic writing, behavioral engagement

Introduction

Educational institutes take different measures to improve their graduate students' academic writing ability. These supportive steps include academic writing and research courses, instructor feedback, peer feedback, group writing sessions, and automated feedback in recent years. These activities can help inexperienced writers boost their academic writing ability and enable them to share their research findings in different output types, including journals, conferences, monographs, etc.

Although these supportive measures have been tried out in practice and research contexts, little is known about how doctoral students engage behaviorally with peer and automated feedback on their dissertation proposals. The examination of these two innovative pedagogical designs can show how effective they are in a graduate-level academic writing context (i.e., dissertation proposal writing). The study of the literature shows that the investigation of students' behavioral engagement with these two feedback types in an academic context has remained unexplored. Thus, the current research aimed to investigate Teaching English as a Foreign Language doctoral students' behavioral engagement with peer and automated feedback on their dissertation proposals.

Literature Review

Peer Feedback on Graduate Students' Proposals

Some prior studies have investigated peer feedback on graduate students' thesis proposals. Here is a brief review of these studies. (Chen, 2010) investigated learners' attitudes toward exchanging peer feedback and receiving teacher comments in a postgraduate degree. The results of his cross-sectional study showed that the respondents were mostly positive about exchanging comments; however, they were careful about language-focused peer comments. Furthermore, the graduate students found teachers' comments beneficial, but the usefulness varied in various contexts. (Saeed & Ghazali, 2019) also studied how graduate students compose their academic texts and implement comments on their texts. The study of the data shows that the research group was interested in developing their knowledge through peer feedback practices, how to compose and present proposals, and finding directions in research proposals.

Yu et al. (2020) carried out research to check master's degree students with peer feedback at a Macau university. She used a complete package of data, including face-to-face semi-structured interviews, stimulated recalls, online interviews, different drafts of master's theses, peer-written comments, audio recordings of oral peer feedback conferences, and the last version of master's theses to answer the research questions. The results indicated that the three engagement types were related and affected each other significantly; however, the most significant association was found between emotional engagement and behavioral and cognitive engagement types. Similarly, (Yu, 2019) investigated the extent to which providing peer feedback can improve graduate students' academic writing ability. The findings of her study showed that providing peer feedback on peers' texts improved their academic writing skills and raised their awareness of the thesis genre. In this study, the participants became more strategic learners by seeking external assistance and more reflective academic writers who could write academic texts carefully.

Finally, Al Qunayeer (2020) also examined the opportunities and challenges of using peer feedback in the proposal composition process in a Malaysian context. The results of this study indicated that postgraduate students had a positive attitude toward peer feedback; however, they did not have active participation at the beginning of the project. Some students did not identify peer comments as dependable. In some other cases, graduate students were worried about sharing their ideas with others since they thought someone else might use them.

Automated Feedback

The second independent variable of this study is automated feedback. While word processing applications such as Microsoft Word have been providing basic proofreading services for more than twenty years, several academic and commercial applications have been provided to provide comments on general and academic aspects of writing. This collection includes *E-Rater* and *IntelliMetric*, which have been developed by the Educational Testing Service (ETS) to assign scores to high-stakes tests such as TOEFL (Test of English as a Foreign Language) or GRE (Graduate Record Examination). E-rater examines the texts by analyzing lexical complexity, syntactic accuracy, mechanics, stylistic features, organization, and idiomatic expressions (Burstein et al., 2004). Similarly, *IntelliMetric* examines the texts based on focus and unity,

development of content, organization, sentence structure, mechanics and conventions, and latent semantic dimensions. Another sophisticated platform, namely *Writing Pal*, has been devised specially for argumentative texts by providing feedback on cohesion, rhetorical style, language use, and linguistic sophistication (McNamara et al., 2013). Maybe, the most well-known online feedback platform is Grammarly, which provides feedback on both the erroneous sections and those areas that can be improved. This commercially successful platform provides feedback on grammar, spelling and usage, wordiness, style, punctuation, tone, and even plagiarism.

In recent years, some attempts have been made to create genre-based feedback-providing platforms, e.g., Research Writing Tutor (Cotos, 2016) and AcaWriter (Knight et al., 2020); however, they have not been accessible to researchers or end-users to examine how efficient they are in providing a wide range of comments to learners. The main disadvantage of these platforms was their mere focus on the rhetorical structure and no other criteria. Thus, although these platforms were useful in different respects, none of them could provide feedback on all aspects of academic writing. Over the last two years, ChatGPT 3.5 has been used by individuals to get written feedback on texts.

However, the review of the literature revealed that while these studies have investigated the issue of peer feedback on master's degree thesis proposals, which is a significant genre in educational settings, none of the prior studies have investigated learners' engagement with peer feedback on Ph.D. dissertation proposals. The present research aimed to occupy these niches in the literature and investigated how doctoral students behaviorally engaged with peer and automated feedback on their dissertation proposals. The following research questions guided this study:

RQ 1: What feedback types are provided by ChatGPT 3.5 and Ph.D. students on L2 dissertation proposals?

RQ 2: How do Ph.D. students engage with peer ChatGPT-generated feedback on their dissertation proposals behaviorally?

Method

Participants

Twenty-eight Teaching English as a Foreign Language (TEFL) Ph.D. students from five state and Islamic Azad universities participated in this study. These students had passed their comprehensive exams and were writing their dissertation proposals. The participants were selected based on convenience sampling; however, the researchers selected the samples from three different provinces in Iran (Fars, Tehran, as well as Kohgiule and Boirahmad) to have a more representative sample. The participants were native speakers of Persian, and their ages ranged between 28 and 38 (M = 31.5, SD = 2.7). The self-reported English language proficiency of the participants showed that they were B2 (n = 8), C1 (n = 14), and C2 (n = 6) English language users. The participants included both male (n = 16) and female (n = 12) students. The participation in this study was voluntary, and no monetary incentive was employed to encourage the participants to take part in the present research.

Documents

The researcher collected three sets of documents in this research. First, the participants' first draft of their proposals. The dissertation proposals with different research approaches (i.e., quantitative (N = 13), qualitative (N = 6), and mixed-methods (N = 9) were included in this research. Second, the comments provided by the Ph.D. students and the automated feedback platform were also recorded. The third data set used was the revised drafts of the proposals mentioned earlier. The participants were asked to revise their texts in two weeks.

Instrument and Materials

Stimulated-Recall Interview

The researcher also employed stimulated recall interviews to identify the participants' reasons for not implementing the comments provided by their peers and the automated feedback platform. To do so, the researcher used the first draft of the texts, the second draft, and the comments as the recall materials. The interviews were in the participants' mother tongue (Farsi). The interviews were conducted both online (n = 11) and in-person (n = 17) based on their own preference. Since stimulated recall interviews could be a new activity for some of the participants, the researcher asked them to provide a simple description of a photo and solve a simple multiplication to practice how stimulated recall interviews work. The interviews were audio recorded and transcribed for further analysis.

Automated Feedback Platform

The researcher employed ChatGPT 3.5 to provide feedback on the participants' texts. The texts were inserted into the platform, and the researcher provided the platform with five different prompts. One of the prompts was "Provide detailed feedback on the grammatical and punctuation aspects of this text". Similar prompts were also inserted to examine the text organization (rhetorical moves), content, formatting (APA 7), word choice, and coherence and cohesion.

Data Collection Procedure

The data collection in this research took seven months. The first step in the data collection was obtaining written consent from the participants of the present study. Then, the researcher randomly assigned the participants to two groups: peer feedback and automated feedback. The participants were asked to write their proposals and send them to the researcher through email. The researcher anonymized the texts and sent the proposals of those who were in the peer feedback group to the Ph.D. students who participated in the study. In the automated feedback group, the researcher inserted the texts into ChatGPT 3.5 and used a set of prompts to elicit comments. The feedback receivers in both groups had two weeks to revise their texts based on the comments and send both versions to their supervisor and the researcher. This study started with 34 PhD students, but six participants dropped out during the data collection process. The researcher asked ChatGPT and PhD students to provide feedback based on the same criteria of grammar and punctuation, text organization (rhetorical moves), content, formatting (e.g., APA 7), word choice, and coherence and cohesion.

Data Analysis

In order to categorize the comments provided by ChatGPT 3.5 and Iranian L2 doctoral students on doctoral students' dissertation proposals, the researcher used two categorizations. The first categorization was inspired by the oft-cited analytic writing scoring scheme by (Jacobs, 1981). While the researchers categorized the files inductively, they used Jacobs et al.'s model to label the identified categories. To ensure the accuracy of this categorization, the first author categorized the comments, and an applied linguistics PhD-holder categorized half of the comments separately. The consistency of the categorization done by the independent coders was computed ($r = .92$). Then, the coders discussed the discrepancies until they reached the same decisions.

In addition, the researchers used the categorization provided by Berndt et al. (2018) to label the comments based on their specificity: general feedback (providing only the faulty area and its issues), elaborated feedback, identifying the erroneous area and providing guidance on how to fix the problem, and justified feedback, providing information on why the item is erroneous and why the corrected version or the suggestion is a better choice. The researchers of this study put the comments into these three classes deductively, and a consistency level of .96 was obtained.

To examine whether the comments were applied in the revised versions or not, the researchers studied both the first and revised versions. In those cases, where the researchers could not make sure that the comments were applied or not, the proposal writers were asked if they had applied the intended comments. The findings of the comparison of the first author and an applied linguistics PhD-holder out of the research team showed a high consistency level (r = .97).

Finally, to identify the participants' reasons for leaving some comments unincorporated, the researchers used stimulated recall interviews. The researchers employed thematic analysis to analyze the collected interview data. Thematic analysis procedure steps (i.e., familiarization, open coding, closed coding, and thematic categorization) were used to analyze the collected data. The reasons mentioned by the participants were categorized thematically and the frequency of each item was computed. To ensure the accuracy of this analysis, the second researcher examined half of the data deductively (based on the categories identified by the first author), and a high level of consistency (r = .97) was achieved.

Results

Feedback Types

Both PhD students and ChatGPT 3.5 were asked to provide feedback on grammar and punctuation, text organization (rhetorical moves), content, formatting (APA 7), word choice, and coherence and cohesion. Table 1 provides a summary of the feedback types provided by these two feedback sources.

Table 1

Feedback Types Provided by ChatGPT 3.5 and Peers

	ChatGPT 3.5	Peer feedback
Grammar	731 (M = 26.10, SD = 1.9)	403 (M = 14.39, SD = 2.3)
Organization	664 (M = 23.71, SD = 2.13)	286 (M = 10.21, SD = 1.17)
Content	436 (M = 15.57, SD = 1.78)	472 (M = 16.86, SD = 2.3)
Formatting	562 (M = 20.07, SD = 1.3)	388 (M = 13.85, SD = 1.7)
Word choice	319 (M = 11.39, SD = .89)	207 (M = 7.39, SD = 1.2)
Total	2712 (M = 96.85, SD = 6.42)	1756 (M = 62.71, SD = 7.8)

As Table 1 shows, ChatGPT provided 731 comments on grammatical and punctuation mistakes or areas that could be improved (M = 26.10, SD = 1.9) while students provided 403 comments (M = 14.39, SD = 2.3) on these areas. The second most frequent feedback type provided by ChatGPT was organization (N = 664, M = 23.71, SD = 2.13); however, significantly fewer peer comments were provided on this area (N = 286, M = 10.21, SD = 1.17). Formatting was the third most frequent feedback type by the implemented generative AI tool (N = 562, M = 20.07, SD = 1.3), and 388 comments (M = 13.85, SD = 1.7) were given by the students on their peers' texts. Content was the next feedback type that was provided by ChatGPT (N = 436, M = 15.57, SD = 1.78) and PhD students (N = 472, M = 16.86, SD = 2.3). While the difference between the numbers of comments provided by these two sources was not significantly different, it was the only area that PhD students provided more feedback on than the generative AI platform used in this study. Finally, the lowest number of comments was provided on word choice by ChatGPT (N = 319, M = 11.39, SD = .89) and PhD students (N = 207, M = 7.39, SD = 1.2).

The researchers also categorized the comments based on Berndt et al.'s (2018) model. In this categorization, comments are put into three classes of general comments, elaborated comments, and justified comments. The findings of this analysis are provided in Table 2.

Table 2

Feedback Provided by ChatGPT and Peers based on the Level of Specificity

	ChatGPT feedback		Peer feedback	
Level of specificity	Frequency	Percentage	Frequency	Percentage
General feedback	0	0	462	26.3
Elaborated feedback	891	32.85	912	51.93
Justified feedback	1821	67.14	382	21.75
Total	2712	100	1756	100

As given in Table 2, since the prompt given to ChatGPT asked this platform to provide justified feedback, all comments were justified (N = 2712, 100 %); however, the comments provided by the students were of different levels of specificity. Just over a quarter of the comments were general ones (N = 462, 26.3 %). Around half of the comments (N = 912, 51.93 %) included elaborated information on the areas identified as erroneous. Finally, around one-fifth of the comments provided by the PhD students were justified (N = 382, 21.75 %).

Feedback Incorporation

The second part of this study was the students' levels of applying peer feedback and GenAI-generated feedback. To calculate the extent to which the participants applied comments received from these two feedback sources, the researcher compared the students' first drafts and second drafts carefully and labeled comments as applied, text changed, and overlooked. Table 3 provides a summary of the findings pertinent to this part of the research.

Table 3

ChatGPT-Generated Feedback Incorporation Rate

	Applied	Text modified	Ignored	Total
Elaborated feedback	577 (64.75 %)	72 (8.08 %)	242 (27.16 %)	891
Justified feedback	1496 (82.15 %)	107 (5.87 %)	218 (11.97 %)	1821
Overall	2073 (76.43 %)	179 (6.6 %)	460 (16.96 %)	2712

As presented in Table 3, the students applied 577 (64.75 %) of the elaborated comments provided by the GenAI tool used in this research. In reaction to a modest 8.08 percent of elaborated comments (N = 72), the students modified their texts, and around a quarter of the comments (N = 242, 27.16 %) were ignored by the PhD students. Regarding justified comments, the participants applied 1496 of the GenAI-generated comments (82.15 %), and around one-tenth of the comments (N = 218, 11.97 %) were overlooked by the students. They also modified their texts in response to 107 comments (5.87 %).

Table 4

Peer Feedback Incorporation Rate

	Applied	Text modified	Ignored	Total
General feedback	268 (58 %)	16 (3.46 %)	178 (38.52 %)	462
Elaborated feedback	576 (63.15 %)	153 (16.77 %)	183 (20.06 %)	912
Justified feedback	286 (74.86 %)	33 (8.63 %)	63 (16.49 %)	382
Total	1130 (64.35 %)	202 (11.5 %)	424 (24.14 %)	1756

As indicated in Table 4, the lowest percentage of feedback incorporation belonged to general comments. More than half of these comments (N = 268, 58 %) were applied by the participants. The PhD students ignored more than one-third of general comments (N = 178, 38.52 %), and they modified their own texts in response to general feedback in 16 cases (3.46 %). The results also showed that elaborated comments were applied in 63.15 percent of cases (N = 576), and one-fifth of these comments (N = 183) were ignored by the participants. The participants modified their texts as a reaction to 153 elaborated comments (16.77 %). Finally, the justified comments had the highest level of incorporation. Around 75 percent of justified comments (74.86 %) were applied by the participants of this study, and only 16.49 percent of the comments were ignored by the PhD students participating in this study. The students also changed their texts in 33 cases (8.63 %) and did not apply the comments. Overall, the students applied around two-thirds of the peer comments were applied by the PhD students, and 24.14 percent of the comments were ignored by the participants.

These students also changed their texts in response to the provided comments in 202 cases (N = 11.5 %).

Reasons for Leaving Comments Unincorporated

The participants of this study left some comments unincorporated. By unincorporated, we mean those comments that have been ignored by the feedback receivers or those comments that resulted in the deletion or modification of a section, ranging from a word to a whole paragraph, to avoid applying the comments. Four main reasons were mentioned by the participants of the current study. The first one was the broad nature of comments that could not guide students to modify the text. The students did not apply some of the comments for not being accurate or relevant. The third reason mentioned by the participants was the high level (beyond learners' perceived ability) of requirements embedded in comments. The last reason identified in the stimulated recall interviews was the students' difficulty understanding the comments provided by ChatGPT or peers. Table 5 provides a report of the frequencies of these reasons for the comments given by ChatGPT and doctoral students.

Table 5

Reasons for Leaving Comments Unincorporated

	ChatGPT 3.5				Peer feedback			
	Broad	Inacc	difficult to apply	Difficult to understand	Broad	Inacc	difficult to apply	Difficult to understand
Grammar and punctuation	14	27	7	13	69	26	34	43
Organization	23	30	6	11	54	58	28	23
Content	37	317	63	23	123	233	46	71
Formatting	6	15	4	7	49	24	16	21
Word choice	3	28	6	3	56	13	4	27
Total	83 (13 %)	413 (64.63 %)	86 (13.45 %)	57 (8.92 %)	294 (47.0 %)	134 (21.4 %)	116 (18.53 %)	82 (13.1 %)

As indicated in Table 5, different reasons were mentioned by the participants for not applying the comments. The analysis of the data showed that the comments were not applied since they were too broad, inaccurate, beyond the feedback receivers' ability, and difficult to comprehend. The scrutiny of the data showed that the majority of GenAI-generated comments were not applied for being inaccurate (N = 413, 64.45 %). The participants (N = 86, 13.45 %) also mentioned that the requirements of the comments were difficult to apply. Thirteen percent of the comments (N = 83) were not applied for being too broad. Finally, less than ten percent of the comments provided by the Gen-AI platform (N = 57, 8.92 %) were left unincorporated as they were difficult for the participants to understand.

The analysis of the data also revealed that Iranian doctoral students did not apply around half of the peer comments (N = 294, 47 %) for being too broad. The students left 134 comments (18.53 %) unincorporated for being inaccurate. The third most frequent reason for the unincorporated comments provided by students was the requirements of the comments which feedback receivers believed were beyond their ability (N = 116, 18.53 %). Finally, the participants did not apply 82 comments (13.1 %) as they found them difficult to understand.

Discussion

The first question of this study addressed the feedback types provided by ChatGPT 3.5 and peers on their dissertation proposals. The findings revealed that ChatGPT 3.5 was successful in providing both form-related and content-related aspects of the texts. ChatGPT 3.5 proved capable of examining the content of a complicated academic text and providing feedback on the content and the way the arguments should be organized. This is in line with the findings of prior studies (Awidi, 2024; Steiss et al., 2024) which have shown that GenAI can be a relatively successful tool for providing feedback on a wide range of aspects of written products. While these studies were conducted in argumentative essay-writing contexts, the present study contributed to the literature by providing evidence for the affordances of ChatGPT 3.5, as a GenAI platform, to analyze extended academic texts such as dissertation proposals and provide feedback on different aspects.

The analysis of the comments given by the PhD students showed that they provided comments on both content-related and form-related aspects of the proposals. These novice researchers could analyze their peers' proposals and provide feedback on both global and local elements. This finding is in line with some studies in the literature (Chen, 2010; Yu et al., 2020) that showed graduate-level students' ability to provide feedback on different aspects. Although these studies were conducted in different contexts (e.g., class writing assignments and master's theses), the collection of these studies shows the capability of graduate-level students to provide high-quality peer feedback.

Another finding of the study dealt with the level of specificity of the comments provided by the GenAI tool and PhD students. While ChatGPT 3.5 was successful in providing elaborated and justified feedback and no general feedback was provided, a quarter of the comments given by PhD students were general. The main reasons for ChatGPT's success in providing specific feedback can be its technological power and the quality of the prompt that the researcher used to elicit favorable results. Previous studies (Law, 2024; Octavio et al., 2024) have underlined the significance of using suitable prompts that can result in high-quality data. Thus, the suitability of the prompt that required the GenAI tool (i.e., ChatGPT 3.5) to provide elaborated and justified feedback seems to be effective.

On the other hand, in this research, the PhD students were asked to provide elaborated and justified feedback on their peers' texts. The findings showed that the PhD students in this study gave general comments in only a quarter of the cases. This study showed that even in the examined context where the participants were only provided with a couple of samples for preferred feedback types (elaborated and justified), they managed to provide more specific comments. Again, one of the reasons for this performance can be doctoral students' ability to provide explanations and justifications within their comments. Previous studies (Berndt et al., 2018; Bolzer et al., 2015) have also shown that high-level students are more likely to provide specific comments on their peers' texts.

The scrutiny of the incorporation pattern showed that the students applied most of the comments and that the specificity of the comments could affect the incorporation level done by the PhD students participating in this study. The findings of this study supported the findings of the study by Mehrpour et al. (2023) that peer comments are given at different specificity levels. Most of the

comments in the present research were elaborated feedback, identifying the erroneous area and providing guidance on how to fix the problem. General feedback, providing only the faulty area and its issues, was the second most frequent type. This type of feedback is usually so brief that it can assist learners in revising their texts minimally because it is likely to be difficult for them to understand how to change their texts to incorporate these comments. The least frequent feedback type was justified feedback. Less than a quarter of the comments were of this type, which not only provides the area and the type of the erroneous items but also gives an explanation of why the provided change should be made to improve the text and how it should be done.

It seems logical that doctoral students provide a wide range of feedback types on their peers' dissertation proposals since they have to analyze a wide range of elements, including both high-order and low-order writing components (Suzuki et al., 2019). Moreover, Pearson (2022) argued that feedback providers address different aspects for various reasons. Some might give general feedback because they believe the recipients can easily apply the comments, so they do not provide detailed information about the errors. Alternatively, feedback providers may recognize an issue but lack the knowledge to guide their peers effectively. Consequently, Although the participants were asked to provide detailed and justified feedback, they gave feedback with varying levels of specificity based on the nature of the errors and their own understanding and perceptions of the context. Overall, the examination of the provided comments showed that both ChatGPT 3.5 and doctoral students were capable of providing comprehensive (i.e., including different local and global aspects) and specific comments. The examination of the way students engage cognitively can give us further information about the quality of these two feedback types.

The findings also revealed that L2 doctoral students applied 76 percent of the ChatGPT-generated comments and 65 percent of their peers' comments. These numbers denote high levels of behavioral engagement. The analysis of stimulated recall interviews revealed four main reasons for the unincorporated comments. The participants did not apply comments since they were too broad, inaccurate, difficult to apply, and difficult to understand.

The analysis of the data also showed that feedback content in terms of specificity could noticeably affect doctoral students' engagement with comments. Those comments that included justification in addition to a detailed

correction could engage students more than those comments that were general. These findings were also witnessed in previous studies in which learners were reported to engage with specific comments (Fernando, 2020; Wu & Schunn, 2020).

In general, incorporating comments is cognitively demanding (Bitchener, 2017), and general comments can increase the cognitive load of feedback incorporation tasks (Wu & Schunn, 2020) since L2 learners must navigate the complex task of identifying the requirements of these comments (Lachner & Neuburg, 2019). Furthermore, general comments can also negatively impact learners' emotions because they can cause higher anxiety levels due to the lack of specific instructions on how to proceed (Fernando, 2020). This feeling of uncertainty can cause negative feelings and can complicate the situation by disrupting the cognitive and behavioral engagement of L2 writers. This uncertainty may lead to decreased task self-confidence and motivation (Stevenson & Phakiti, 2019). This reduction in self-confidence and motivation can subsequently affect the behavioral and cognitive engagement with other comments in the same and future feedback incorporation tasks.

The comparison of the numbers of ChatGPT-generated and peer comments which were not applied due to their broad nature suggested that the GenAI platform was more successful than peers in providing specific comments. The reason might stem from the quality of the prompt the researcher used to elicit feedback in the ChatGPT condition. When the prompt carefully asks for specific comments, feedback receivers likely get specific feedback. The importance of the quality of prompts in educational activities aided by GenAI tools has been emphasized in previous studies (Law, 2024; Octavio et al., 2024). It appears that the carefully crafted prompts used in this research led to a low level of unincorporated comments for being too broad.

Doctoral degree students also stated that some comments were difficult to understand. They argued that these comments were beyond their ability, so they had no choice but to ignore them. Previous studies have shown the negative effects of mismatched comments and learners' knowledge (Davin, 2013). Since the 1980s, the suitability of comments relative to learners' (self-perceived) levels has been controversial. Even in approaches such as sociocultural theory, where feedback is crucial for learning, the issue of reciprocity, how learners respond to mediation, often provided as feedback (Poehner & Wang, 2021), plays a

significant role in the success of feedback activities. This responsiveness, reflecting learners' engagement with feedback, is important because comments do not alter learners' cognitive structures if they have not reached the required cognitive ability.

In the present study, the participants reported that some comments were beyond their ability. However, in both ChatGPT-generated and peer feedback cases, this reason did not go beyond 18 percent of the unincorporated comments. It seems that both feedback types included explicit or implicit requirements that were manageable for doctoral students in most cases. The first reason that can be mentioned for both conditions is that feedback receivers in this study were doctoral students who were capable of modifying their texts based on the comments since they were familiar with the standards and requirements of an acceptable academic text. This knowledge gave them the perception that they could apply the overwhelming majority of the comments. The second reason is related to peer feedback and addresses the students' familiarity with their peers' knowledge, weaknesses, and abilities. Prior studies (Vuogan & Li, 2023; Yu & Lee, 2016) have shown that students, due to their interactions with their peers were able to give comments that were not beyond their peers' ability in most cases.

The third issue that the participants mentioned for not applying comments was the inaccuracies that penetrated the comments. The analysis of peer feedback literature reveals reservations about its accuracy. Prior studies have shown that inaccurate peer comments can negatively affect learners' perceptions of peer feedback (Van der Kleij & Lipnevich, 2021). This can result in decreased engagement with comments due to uncertainty about their accuracy (Sluijsmans et al., 2002). Trust has been identified as a significant factor in students' engagement with feedback (Sedikides et al., 2016). The literature indicates that students are less trustful when comments come from a perceived less competent peer (Zhai & Ma, 2023). The presence of this theme in the present data indicates that even in doctoral-level writing contexts, believing in the accuracy of the provided comments is an influential factor that should not be overlooked.

Participants in the present study frequently mentioned the inaccuracy of ChatGPT-generated comments. Approximately two-thirds of the unincorporated comments provided by GenAI were flagged as inaccurate by the students. In this

study, inaccuracy refers to both false and irrelevant information, with around twenty percent falling into the latter category. This is in line with the findings of recent studies on GenAI-generated feedback, which have found inaccuracy as a main drawback of materials produced by GenAI (Wang et al., 2024). As Wang et al. (2024) state, while GenAI tools can provide substantial amounts of useful information, faulty information can sneak into the results. This necessitates students' careful use of the received information. The findings of the present study reveal that doctoral students examined the comments provided by ChatGPT, enabling them to identify a significant number of inaccuracies. Their ability to detect inaccuracies may be attributed to doctoral students' knowledge and the significance they assign to their texts (i.e., dissertation proposals) could possibly motivate them to scrutinize all comments meticulously. However, this level of scrutiny may not occur when the texts are not related to high-stakes conditions or when the students are less competent (e.g., master's degree or undergraduate students).

The fourth factor that was mentioned by L2 doctoral students was the incomprehensibility of comments. Incorporating a comment is not possible if feedback receivers cannot understand the comments provided on their texts (Fan & Xu, 2020; Han, 2017). Prior research has shown that a disadvantage of written feedback is that feedback providers cannot gauge the extent to which their comments are understood until they review the revised version (Ellis, 2010). Moreover, feedback receivers cannot immediately ask for clarification when the feedback provider is not accessible. This chronological gap may lead to an inability to understand or a misunderstanding of comments, which disrupts the feedback incorporation process.

Sachs and Polio (2007) emphasizing the importance of feedback understanding, argue that ensuring feedback receivers' identification and comprehension is crucial for effective feedback uptake. Examination of dual-layered awareness in feedback activities has shown that mere noticing is insufficient; students must achieve a level of understanding to benefit from comments (Rosa & Leow, 2004). To move beyond superficial awareness, teachers should ensure students' comprehension (Han & Hyland, 2015). In the present study, the issue of incomprehensible comments was not a major one since around 10 and 13 percent of the ignored ChatGPT-generated and peer feedback were labeled difficult to understand. Thus, although this theme

emerged in the results, it does not seem to be a serious problem for doctoral students in either GenAI or peer feedback contexts.

The findings of the present study provided further empirical evidence for the significance of providing elaborated and justified feedback. In line with previous studies in the literature (Berndt et al., 2018; Bolzer et al., 2015), justified comments were applied more than elaborated ones in the present context. These results suggest that to have higher levels of behavioral engagement by doctoral students, feedback needs to be elaborated and justified. However, although behavioral engagement and cognitive engagement are related, further longitudinal studies are required to investigate whether and the extent to which elaborated and justified feedback can improve students' academic writing ability in the long run.

According to the findings, the main reason for unincorporated comments provided by ChatGPT 3.5 is the accuracy of the comments. Although 85 percent of the comments were identified as accurate by feedback receivers in this study, a modest 15 percent of the comments were inaccurate. It seems that while ChatGPT 3.5 can be regarded as a highly reliable source of feedback on academic texts, learners need to be cautious about the accuracy of the comments provided by this GenAI platform. The identification of these inaccurate comments suggests that doctoral students are aware of this possible drawback of GenAI-generated comments and flag 15 percent of these comments as inaccurate. Considering these findings, policymakers, supervisors, and instructors who intend to integrate GenAI-generated feedback into academic writing contexts should ensure the students' AI literacy for academic writing purposes. This issue that has been emphasized in previous theoretical and empirical studies (Wang et al., 2024) suggests that boot camp workshops can be held to make sure that graduate-level students are capable of using the comments provided on their academic texts by GenAI platforms efficiently.

Conflict of interest: None

References

Al Qunayeer, H. S. (2020). Supporting postgraduates in research proposals through peer feedback in a Malaysian university. *Journal of Further and Higher Education*, *44*(7), 956-970.

Awidi, I. T. (2024). Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence*, *6*, 100226.

Berndt, M., Strijbos, J.-W., & Fischer, F. (2018). Effects of written peer-feedback content and sender's competence on perceptions, performance, and mindful cognitive processing. *European Journal of Psychology of Education*, *33*, 31-49.

Bitchener, J. (2017). Why some L2 learners fail to benefit from written corrective feedback. In *Corrective feedback in second language teaching and learning* (pp. 129-140). Routledge.

Bolzer, M., Strijbos, J.-W., & Fischer, F. (2015). Inferring mindful cognitive-processing of peer- feedback via eye- tracking: Role of feedback-characteristics, fixation- durations and transitions. *Journal of Computer Assisted Learning*, *31*(5), 422-434.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *Ai magazine*, *25*(3), 27-27.

Chen, C. W.-y. (2010). Graduate students' self-reported perspectives regarding peer feedback and feedback from writing consultants. *Asia Pacific Education Review*, *11*, 151-158.

Cotos, E. (2016). Computer-assisted research writing in the disciplines. In *Adaptive educational technologies for literacy instruction* (pp. 225-242). Routledge.

Davin, K. J. (2013). Integration of dynamic assessment and instructional conversations to promote development and improve assessment in the language classroom. *Language Teaching Research*, *17*(3), 303-322.

Ellis, R. (2010). Epilogue: A framework for investigating oral and written corrective feedback. *Studies in second language acquisition*, *32*(2), 335-349.

Fan, Y., & Xu, J. (2020). Exploring student engagement with peer feedback on L2 writing. *Journal of Second Language Writing*, *50*, 100775.

Fernando, W. (2020). Moodle quizzes and their usability for formative assessment of academic writing. *Assessing Writing*, *46*, 100485.

Han, Y. (2017). Mediating and being mediated: Learner beliefs and learner engagement with written corrective feedback. *System*, *69*, 133-142.

Han, Y., & Hyland, F. (2015). Exploring learner engagement with written corrective feedback in a Chinese tertiary EFL classroom. *Journal of second language writing*, *30*, 31-44.

Jacobs, H. L. (1981). *Testing ESL composition: A practical approach. English composition program*. ERIC.

Knight, S., Shibani, A., Abel, S., Gibson, A., & Ryan, P. (2020). AcaWriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research*.

Lachner, A., & Neuburg, C. (2019). Learning by writing explanations: Computer-based feedback about the explanatory cohesion enhances students' transfer. *Instructional Science*, *47*(1), 19-37.

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, 100174.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*, 499-515.

Mehrpour, S., Hoomanfard, M. H., & Vazin, E. (2023). Peer Feedback Accuracy in Synchronous and Asynchronous Computer-mediated Conditions in an EFL Context. *Iranian Journal of Language Teaching Research*, *11*(1), 97-116.

Octavio, M. M., Argüello, M. V. G., & Pujolà, J.-T. (2024). ChatGPT as an AI L2 teaching support: A case study of an EFL teacher. *Technology in Language Teaching & Learning*, *6*(1), 1-25.

Pearson, W. S. (2022). Response to written commentary in preparation for high-stakes second language writing assessment. *Asian-Pacific Journal of Second and Foreign Language Education*, *7*(1), 19.

Poehner, M. E., & Wang, Z. (2021). Dynamic assessment and second language development. *Language Teaching*, *54*(4), 472-490.

Rosa, E. M., & Leow, R. P. (2004). Awareness, different learning conditions, and second language development. *Applied psycholinguistics*, *25*(2), 269-292.

Sachs, R., & Polio, C. (2007). Learners' uses of two types of written feedback on a L2 writing revision task. *Studies in Second Language Acquisition*, *29*(1), 67-100.

Saeed, M. A., & Ghazali, K. (2019). Engaging postgraduates in a peer research group at the research proposal stage in a Malaysian university: support and challenges. *Teaching in Higher Education*, *24*(2), 180-196.

Sedikides, C., Luke, M. A., & Hepper, E. G. (2016). Enhancing feedback and improving feedback: Subjective perceptions, psychological consequences, behavioral outcomes. *Journal of Applied Social Psychology*, *46*(12), 687-700.

Sluijsmans, D. M., Brand-Gruwel, S., van Merriënboer, J. J., & Bastiaens, T. J. (2002). The training of peer assessment skills to promote the development of reflection skills in teacher education. *Studies in Educational Evaluation*, *29*(1), 23-42.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J.,…Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 101894.

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

Suzuki, W., Nassaji, H., & Sato, K. (2019). The effects of feedback explicitness and type of target structure on accuracy in revision and new pieces of writing. *System*, *81*, 135-145.

Van der Kleij, F. M., & Lipnevich, A. A. (2021). Student perceptions of assessment feedback: A critical scoping review and call for research. *Educational assessment, evaluation and accountability*, *33*, 345-373.

Vuogan, A., & Li, S. (2023). Examining the effectiveness of peer feedback in second language writing: A meta- analysis. *Tesol Quarterly*, *57*(4), 1115-1138.

Wang, N., Wang, X., & Su, Y.-S. (2024). Critical analysis of the technological affordances, challenges and future directions of Generative AI in education: a systematic review. *Asia Pacific Journal of Education*, *44*(1), 139-155.

Wu, Y., & Schunn, C. D. (2020). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology*, *60*, 101826.

Yu, S. (2019). Learning from giving peer feedback on postgraduate theses: Voices from master's students in the Macau EFL context. *Assessing Writing*, *40*, 42-52.

Yu, S., Jiang, L., & Zhou, N. (2020). Investigating what feedback practices contribute to students' writing motivation and engagement in Chinese EFL context: A large scale study. *Assessing Writing*, *44*, 100451.

Yu, S., & Lee, I. (2016). Peer feedback in second language writing (2005–2014). *Language Teaching*, *49*(4), 461-493.

Zhai, N., & Ma, X. (2023). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *Journal of Educational Computing Research*, *61*(4), 875-900.

Biodata

Lotfollah Akbarpour is a PhD student at the Islamic Azad University, Bushehr Branch. He has been an English teacher for more than 30 years. He has authored the book *Word Formation Practice Workshop* and authored several research papers published in national and international journals.

Leila Saberi is an assistant professor whose work is focused on teaching English to EFL learners. Her research is focused on educational issues and publications are in different domestic and foreign magazines. She has taught different courses in this field to M.A. and Ph. D. students and supervised more than a hundred theses and dissertations. In addition, she has been the head of the English department for more than 10 years.

Saeed Yazdani is an associate professor in English language and literature. He has been teaching for 30 years at the Islamic Azad University, Bushehr Branch, and has published seven books and many research articles.

مشارکت دانشجویان دکترای آموزش زبان انگلیسی در بازخورد گروه همسان و بازخورد خودکار بر پیشنهاده ی رساله ها

این مطالعه به منظور بررسی بازخوردهای ارائه شده توسط ChatGPT 3.5 و دانشجویان دکتری بر پیشنهادههای پایان نامه دکتری و مشارکت رفتاری گیرندگان بازخورد با این دو نوع بازخورد نیز مورد بررسی قرار گرفت. شرکت کنندگان در این پژوهش که بر اساس نمونه گیری در دسترس انتخاب شدند، ۲۸ نفر از دانشجویان دکتری آموزش زبان انگلیسی

به عنوان زبان خارجی از سه استان بودند که پیشنهاده ی پایان نامه خود را به زبان انگلیسی نوشتند. نسخه اول و نسخه های اصلاح شده پیشنهاده آنها و نظرات ارائه شده برای شناسایی انواع بازخورد و میزان اعمال نظرات آنها مورد تجزیه و تحلیل قرار گرفت. علاوه بر این، از مصاحبه های یادآوری برانگیخته شده برای شناسایی دلایل عدم استفاده از برخی از شرکت کنندگان استفاده شد. یافته ها نشان داد که هم ChatGPT 3.5 و هم دانشجویان دکترا در ارائه نظرات مرتبط با محتوا و فرم موفق بودند. این دو منبع بازخورد همچنین تعداد قابل توجهی بازخورد مفصل و با جزییات در مورد پیشنهادات پایان نامه ارائه کردند. گیرندگان بازخورد بیشتر نظرات ارائه شده توسط این منابع بازخورد را اعمال کردند و سطح ویژگی نظرات بر نرخ اعمال تأثیر گذاشت. یافته‌های تجزیه و تحلیل موضوعی داده‌های مصاحبه یادآوری تحریک‌شده نشان داد که شرکت‌کنندگان نظرات را به‌دلیل اینکه بسیار گسترده، نادرست، دشوار برای اعمال و درک دشوار بودند، اعمال نکردند.

کلمات کلیدی: بازخورد ایجاد شده توسط ChatGPT 3.5، بازخورد همسالان، نوشتن تحصیلی، مشارکت رفتاری