# Artificial Intelligence as a Provider of Feedback on EFL Student Compositions

Manal Saleh M. Alghannam[1]

[1] Assistant professor, Department of English Language and literature, College of Languages and Humanities, Qassim University (QU), Buraidah, Qassim region, Saudi Arabia

Correspondence: Manal Saleh M. Alghannam, Assistant professor, Department of English Language and literature1, College of Languages and Humanities, Qassim University (QU), Buraidah, Qassim region, 52571 Saudi Arabia. E-mail: ma.alghannam@qu.edu.sa

https://orcid.org/0000-0002-0124-8417.

## Abstract

In response to the arrival of advanced artificial intelligence (AI) in the form of ChatGPT, this study examines its potential for providing feedback to foreign language writers. This represents a more acceptable use of AI in the writing classroom, rather than students simply using AI to write their entire essay. The methodological procedure involved eliciting normal classroom writing-practice essays from 29 English major students at a Saudi university, with ChatGPT (2023) then given a simple prompt requesting feedback. Both the essays and the feedback were qualitatively analysed to respond to research questions concerning the feedback's consistency and credibility, and the extent to which it represented the different potential feedback types, based on a review of the extensive literature on the subject. Although superficially impressive, close examination revealed certain weaknesses to the AI feedback. For example, there was inconsistency in how the feedback was handled across essays, and some statements were not fully accurate regarding the respective text. In focus, the feedback was primarily accuracy-oriented, while even-handed in attention to content, organisation, and lower-level language matters, providing both positive and negative comments. However, there was a paucity of message-oriented communicative and explicit affective feedback. Like many teachers, ChatGPT was selective in terms of the feedback provided, but the decisions of what to address did not seem altogether motivated by criteria that an expert human feedback provider would consider. The main conclusion is that while AI feedback on writing practice is useful, it does require human monitoring by a teacher.

**Keywords:** AI, ChatGPT, EFL, feedback, writing

## 1. Introduction

Artificial intelligence (AI) has recently made a significant leap forward in terms of its language capability. With the free availability of a state-of-the-art AI facility since late 2022, namely ChatGPT, its potential educational and broader applications are being widely explored. In the domain of written language production, considerable discussion has been generated concerning its possible use by writers, including non-native speakers, in situations where the intention is for the writer to exploit their own ability/proficiency in the language, as opposed an AI-produced outcome (Reich, 2022). The present study does not concern that latter use of AI, widely regarded as *cheating*, although considerable research efforts are being applied to address such improper usage.

Rather, this investigation concerns a less controversial use of AI, namely, as a means to provide feedback that supports students in improving their foreign language writing, which they themselves have produced via their own proficiency in that language. Such feedback is perceived as supporting rather than bypassing learning, and is traditionally expected to be provided by teachers for intermediate or final student drafts, although peer- and self-feedback (via the student's own reflections, although prompted perhaps by a teacher-provided checklist of issues requiring vigilance) is also widely accepted and considered useful (Diab, 2016).

Until recently, computer-based feedback has been unable to provide anything close to teacher-like feedback on extended passages of text written to a set title or prompt, and has otherwise proven very limited in terms of qualitative feedback regarding the content, or the writing approach (beyond what is offered by facilities such as spelling, grammar, and style checkers). Typically only suggestions for low-level corrections have been offered, which are not always appropriate, rather than providing feedback that is highly accurate and embraces broader matters such as the overall text structure and logic of argument (Cotos, 2023). Nevertheless, ChatGPT now promises a more accurate and comprehensive response regarding the levels and types of feedback provided. The present study therefore seeks to examine this phenomenon via a trial of ChatGPT as provider of feedback for English as foreign language (EFL) undergraduate writers.

## 2. Literature Review

This review concerns two areas relevant to the present study: the nature of different feedback types that can be provided for a foreign learner's writing, and the history of computerised attempts to provide such feedback.

*2.1 Writing Feedback*

There is a long and rich literature covering many dimensions of feedback on writing, especially that carried out in secondary and higher education by either first- or second-language learners, often of English (Wirantaka, 2019). However, this domain needs to be investigated afresh in the context of electronic tools that may contribute towards, or fully provide such feedback. This section thus explores what is meant by feedback on EFL writing, and the dimensions it contains.

It should be noted that the present study is focused at the descriptive level of exploring the nature of the feedback that one AI program provides on a learner's writing. Therefore, this review attempts to span the main dimensions on which such feedback can vary, that have been discussed. However, studies and discussions of feedback on writing often overflow into the more evaluative consideration concerning which of these dimensional choices yield superior feedback. That is a far broader and more contested matter, not least due to the existence of differing perspectives on what constitutes effective feedback: that could be feedback which enhances student motivation to write, or which produces improvement in the essay's subsequent draft or which evidences student learning reflected in better future essays (Hyland & Hyland, 2006). Given the scope of the present study, evaluation will only be lightly addressed. Suffice to say that support can be found somewhere for the benefit of almost all the options on all the dimensions of feedback variation, depending on the specific student and writing contexts involved.

2.1.1 Feedback in the EFL Writing Class

According to Scott (2014), "there is no widely accepted definition of feedback" (p. 49). Furthermore, what teachers, students, or indeed AI may perceive as feedback can differ, from that defined by an expert.

In education, however, there is some agreement about the core defining features of feedback. First, it is often described as a form of information about, for example, "reactions to ... a person's performance of a task, etc. which is used as a basis for improvement" (Oxford Languages, 2022), or "any information given to the learner about the learner's performance relative to learning goals. It should aim towards (and be capable of producing) improvement in students' learning" (Education Endowment Foundation, 2022).

Second, those definitions highlight that feedback is perceived as information designed to extend learner development, as opposed to information for its own sake. Notably, Sadler (1989) makes the robust claim that unless the impact of the information on the learner's performance is actually measured, such feedback cannot be regarded as truly constituting feedback at all, but is merely "dangling data" (p. 121). However, most experts (who will be cited here) characterise feedback as the intention of supporting the learner and do not require that an effect is actually delivered (Hyland & Hyland, 2006).

Third, feedback is not merely pedagogical information delivered by a teacher (or other person) with the purpose of benefiting the learner; as implied above, to represent feedback, it must be a response to something recently expressed/carried out by learners (in this case, written production). Therefore, feedback differs from a great deal of teacher-initiated information which is not in response to a learner's production. Nevertheless, an effective teacher will underpin most of the information they present to a class through their insight of that which is already known by the students (York, 2014), and which is ultimately, in a more generally retrospective manner, based on the learners' utterances and actions.

Finally, it is notable that these definitions do not define feedback as assessment, with which feedback can be confused (Didau & Rose, 2016). Indeed, feedback may or may not be associated with assessment; it is often simply a component of day-to-day teaching, loosely termed as *practice* (Mory, 2001) and thus as *formative* in nature. In fact, formal assessments/examinations (especially summative) are often completed without the student receiving any feedback other than an overall mark or grade (Harlen & Deakin Crick, 2002).

A corollary of feedback not necessarily being assessment is that it is not inevitably limited to correctness. Given that feedback is broader, it can encompass the introduction of new related information (e.g., making a comparison or connection), or stating the importance of a topic to be learned, for example (Sales, 1993). However, definitions also exist that restrict feedback to correction, such as "any communication or procedure given to inform a learner of the accuracy of a response, usually to an instructional question" (Mory, 2001, p. 919). Indeed, this narrower corrective feedback (CF) often constitutes the denotation that is tacitly endorsed by teachers, students, and so forth in the classroom (Meunier & Muñoz, 2022).

In the context of the teaching of writing, feedback is in fact often taken simply to be correction (including an indication of what is correct, as well as highlighting mistakes, but often emphasising the latter over the former). Moreover, feedback is commonly further restricted to the correction of language (e.g., grammar, vocabulary, spelling, or syntax) rather than the thesis or content. In second language acquisition (SLA) research, the phrase *focus on forms* has emerged for labelling instruction, including feedback, of this nature (Uysal, 2010). CF can, however, also extend to the meaning or content (Lundgren, 2022) of written or spoken production.

More widely in the literature, there is coverage of the broader range of feedback types that the original definition above accommodates, namely, feedback that is clearly non-corrective and not limited to language. This includes feedback of social and affective types, rather than purely cognitive, which can also help achieve the goal of enhanced learning (Figure 1).
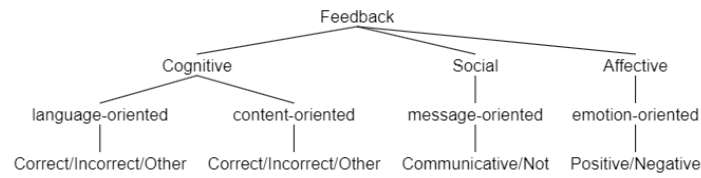
Figure 1. Broad types of feedback on writing

### 2.1.2 The Feedback Provider

Although in an educational setting the most obvious feedback provider is the teacher, there are many others. The feedback types indicated in Figure 1 may be delivered by other adults, such as teaching assistants in the classroom, or friends or parents outside of that environment (El Nokali, Bachman, & Votruba-Drzal, 2010). Moreover, classroom peers are often involved (Ware & O'Dowd, 2008). Furthermore, there is interaction with the types of feedback, whereby the teacher is perhaps the most likely to provide language-oriented feedback while peers, friends, and so forth more typically engage in message-oriented communication.

AI could attempt to fulfil any of these roles, or adopt an additional unique alternative. Historically, *computer-assisted language learning* frequently adopted the proxy teacher role, offering simple kinds of feedback on language correctness, for example, where a writing task effectively had only a small number of correct answers, such as selecting a word or phrase to complete a gap in a text or choosing between four alternative syntaxes for a sentence (Taylor, 1980). Computers and other technology did, however, have the ability to introduce non-verbal feedback, including sounds or visual responses (Nagata, 1993). Later, the computer role was frequently that of a stimulus or resource, neither of which contained a robust feedback component. Currently equipped with its new capabilities, technology is often presented as an assistant or co-*something* for the user (e.g., a copilot or cograder), and the quality and range of possible feedback has improved with the arrival of AI applications such as ChatGPT (see below).

Finally, it has been suggested that the most important provider of feedback to a learner is the learner him/herself, such as when reviewing or editing their written material (Hyland & Hyland, 2006). This is variously referred to as self-feedback, self-assessment, or (when carried out consciously) monitoring. One reason for this importance is that in exam situations, all the other sources of feedback tend to be inaccessible.

In conclusion, Boud and Molloy (2013) and Carless (2020) suggest that rather than conceptualising the teacher as the main source of feedback, they should be considered as the individual who helps facilitate learners to orchestrate and exploit a combination of all the aforementioned feedback sources. Furthermore, this notion is especially advanced in relation to conscious feedback at the tertiary level, a role that AI is also beginning to adopt. For instance, when asked a question, rather than merely offering its own answer, ChatGPT may also suggest that the user seeks information from a certain type of website or individual.

This paper next considers the types of writing feedback in greater detail.

### 2.1.3 Corrective Feedback: Positive and Negative

Feedback on writing in language educational contexts has often been identified with CF, especially *negative* evaluation, or the correction of students' *language* errors (Schachter, 1991). Therefore, it will be of interest to determine whether this type dominates the feedback provided by ChatGPT. Since most teachers and researchers, and indeed students themselves, often hold this view of feedback, the majority of writing tasks tend to be perceived as assessment (resembling tests/exams), rather than an opportunity to engage in judgment-free practice and exploration, thus blurring the important distinction between testing and teaching/learning. In most teaching contexts, this distinction needs to be controlled since despite assessment-like CF's potential for motivational value when carried out occasionally (Devine, 2018), it is often suggested that it should not predominate.

CF need not, however, be limited to the negative correction of language or content, as it can also involve the positive evaluation of accurate language usage or appropriate content. Rather than resembling error correction, this stance is coincidentally closer to following the behaviourist reinforcement paradigm (Lipnevich & Panadero, 2021). At all ages, praise is motivational (see section 2.1.6), while a related concept is the use of the *feedback sandwich* whereby negative comments are bracketed between two elements of positive feedback (Molloy, 2010).

Even correction itself, which superficially appears unavoidably negative, can be presented positively as a *learning step* (Edge, 1989) rather than a deficiency, for example, *Good, but it is better to write xxxx here*. Another ploy is to treat CF, especially of language, as incidental to feedback on other aspects of the written content rather than overtly highlighted consistently. As Edge (1989) notes, feedback on the language usage in EFL classrooms is often foregrounded, rather than that on the coded message (see section 2.1.5). In SLA, this is reflected in the distinction between *focus on forms* and *focus on form* (Uysal, 2010). AI could potentially implement any of the above.

### 2.1.4 Types of Negative Corrective Feedback

Indeed, in mainstream SLA research, the *noticing* of language features is viewed as crucial for learning (Lightbown & Spada, 2006), and of course can be prompted by explicit language-oriented feedback, with the output hypothesis of Swain and Lapkin (1995) drawing

attention to the acquisitional role played by *focus on form* feedback provided on learners' written output. Commonly, however, feedback on small chunks of language, divorced from their role in the overall message of any discourse, termed *focus on forms*, is less frequently endorsed (Uysal, 2010).

Typically, a number of specific types of negative corrective language feedback are recognised, which one might expect to encounter in writing feedback regardless of provider (Panova & Lyster, 2002; Russell, 2009). One type, referred to as *recasting*, involves providing the correct form and thereby only implicitly indicating that there was an error or less than optimal production, and the CF is hence less negative in tone. A second, conversely, is to explicitly indicates that there was an error, but not to provide a correct form; this type includes underlining or circling words in a hard copy essay or writing *E* in the margin, which can prompt the writer to attempt to self-correct. A third type is both to indicate that there is an error, where it can be found, and to provide the correct form (i.e., *explicit correction* through providing full knowledge to inform a correct response); this approach requires the least action from the student. However, explicit correction can acknowledges degrees of errors. Furthermore, there exist partial forms where the feedback provider may merely provide clues, such as inserting a *T* next to a word for a tense error, and *Sp* to indicate a spelling issue. Here, the student is prompted by implication to provide the full correction, which is referred to as *output prompting* (Maolida, 2017). Finally, beyond those approaches there is *elaborated feedback*, which might provide an explanation of the issue or the correction, or perhaps refer back to what had been previously taught in relation to the error.

The correction of content can also fall on a cline. At one end, an actual phrase or sentence is supplied that conveys the correct content, with the striking out of words judged to express erroneous content, or circles and arrows employed to indicate where a word, phrase, or clause should be moved (Russell, 2009). At the other end of the cline is simply underlining or highlighting text where the content is problematic, and perhaps inserting a question mark. In between is the supply of prompts such as *Say why*, or *Does this belong here?*. AI could potentially replicate of these methods.

2.1.5 Communicative Feedback

Within applied linguistic and language teaching research, there have been movements such as the communicative approach, task-based learning, and Krashen's natural approach (Richards & Rodgers, 2001). These have highlighted the importance of message-oriented classroom language use, meaning that the focus, and therefore writing feedback, is not merely on the form or meaning of words and structures, but also focuses on the broader information or opinion that is communicated (the message); this matches the typical function of daily language usage. For example, consider the scenario whereby the students write an essay on their favourite film. Communicative feedback in this context would not be on the correctness of the language (e.g., *The word is spelt character not 'caracter'*), nor the correctness of the content (e.g., *Brad Pitt did not star in the film Titanic*), but rather form a response to the real message being conveyed, namely, that the writer enjoyed the film Titanic (e.g., *I like it too, but I also like the much older film about the same event 'A Night to Remember'*) (Kaye, 2024).

In other words, in communicative feedback as defined here, the feedback provider responds to the writer as a real person might in daily life, albeit that this would usually occur in the first language (L1), such as Arabic in this study. Therefore, the social dynamic momentarily resembles less student-teacher and more of a writer–audience relationship. Chat GPT certainly has the ability to respond in this way.

2.1.6 Emotional/Affective Feedback

Emotional feedback (EF) is definable as that which affects a person's emotional state. EF's separate status was recognised early by Vigil and Oller (1979), and can be clearly distinguished from other feedback, for example, by a feedback provider stating *Good* (positive EF) or *Disappointing* (negative EF). Such examples deliver little information, but will respectively encourage or discourage the writer (Lin, Wu, & Hsueh, 2014). In a computer-assisted environment, a smiling female avatar or smiley face/emoticon can convey such EF without words being required (cf. Terzis, Moridis, & Economides, 2012). Indeed, as Vigil and Oller (1979) indicate, human communication at the affective level is often mediated through kinesics or paralinguistic means rather than language. A human teacher can thus convey considerable EF through their facial expression and tone of voice, accompanied by written feedback (to which the present study is limited) (Paulmann & Weinstein, 2023).

EF (positive or negative) is often delivered alongside CF (Mayordomo, Espasa, & Guasch 2022), such as *Well done, that is correct* or *That is incorrect, pay more attention*. Indeed, CF alone tends to have an unstated affective connotation, whereby an indication of correctness implies positive EF, while an indication of error implies negative EF (Mayordomo et al., 2022) unless the feedback provider counters the inherent negative emotional effect with some explicit, perhaps non-linguistic, positive EF (Vigil & Oller, 1979) such as praising the effort even where accuracy is lacking. The combination of correctness (positive CF) with negative EF is rarely found.

2.1.7 Generality and Completeness

Two other dimensions of the aforementioned feedback (i.e., generality and completeness) are also pertinent (Hyland & Hyland, 2006).

In terms of generality, regardless of the feedback type employed, it can be generalised or global (e.g., for CF, *You need to work on your vocabulary*) or targeted (e.g., *You need to say 'suitable' here*). Generality is often closely aligned with the immediacy of placement (below), with general often located far away, and specific usually close by.

Completeness is also important, since it is rare for a feedback provider to seek to provide feedback that essentially addresses *all* the facets of a piece of writing. In all the aforementioned feedback strategies, a key decision is therefore what to choose to address and what to overlook.

Many criteria may be used in such decision-making. For example, in the present study, the students are English language majors taking a writing class, and therefore the feedback provider might target the language usage rather than the content. Compare this example with the case of essays written by psychology majors taught through the medium of English, where the primary target would likely be the content. Indeed, such a criterion might lead to feedback being limited to the target of the respective lesson (Han, 2002). Another kind of criterion is to tailor the feedback to the needs of each individual in the class. For instance, weaker students might be provided with detailed lower-level language feedback in their known areas of deficiency, alongside more positive affective feedback.

2.1.8 Other Dimensions where Feedback can Vary

Aside from the aforementioned feedback dimensions, a number of others are recognised but will be of less importance to the present study as choices have been made for them as part of the study design, and thus they will not be left for AI to address. Therefore, these feedback dimensions are only briefly discussed below.

**Medium/Channel:** Feedback on writing can be spoken or written, or even take a non-linguistic form (e.g., a tick or an emoji). However, the present case only concerns written feedback, which can be delivered by handwritten annotations, comments in Microsoft Word, email or Facebook, and so forth. In the context of the present study, the feedback would be delivered exactly as produced by ChatGPT, and via email along with a typed version of the original essay text.

**Medium/Language:** Some portions of feedback must be in the target language (in the case of this study, English), namely, the words and structures that are commented on. However, the framing can be in the L1 and not necessarily repeated in the target language, or even be non-verbal (Andersen & Andersen, 2005). Hanif (2020) supports the value of the L1 in the EFL classroom since it decreases the cognitive load and creates positive emotions. However, L1 usage, specifically in feedback, is not often mentioned. Indeed, in lists of functions where the L1 is regarded as useful, feedback is rarely found (e.g., Schweers, 1999). In this study, however, the stance is adopted that for learners at a lower level of English proficiency, the framing of feedback in the L1 could be appropriate. Nevertheless, English is used both for the convenience of the reader here, and also because in the target context (a Saudi university), English usage is expected as the medium of instruction within an English department, although ChatGPT could, without difficulty, render the feedback into Arabic (Modern Standard Arabic) if necessary.

**Audience:** Feedback is typically provided to individual learners separately, and in private (e.g., annotations to their script). However, a teacher may also collate issues from individual students to present to the class (Collin & Quigley, 2021). Nevertheless, the present study is only concerned with the individual audience.

**Immediacy of Placement:** Feedback can be linked directly to something specific produced by the student (e.g., written in the margin with an arrow pointed to the associated text, or in a comment in Microsoft Word). Alternatively, it may be located more distantly at the start or end of a paragraph or the whole text. In the present study, the feedback is aggregated in one place separate from the written text, regardless of its generality, since this is how it is supplied by ChatGPT (Steiss et al., 2024).

**Immediacy of Timing:** Feedback may be delivered directly after the student writes something, and even while they actually writing, or delayed such as at the end of the exercise or lesson, or at the conclusion of the semester (e.g., written term reports). Typically, immediate feedback will be more specific, and delayed feedback more generalised. In addition, immediate feedback (both verbal and non-verbal) more favourably affects student attitudes (Andersen & Andersen, 2005). The broader literature therefore favours immediate feedback, with any delay regarded as placing the learner at a disadvantage through the withholding of pertinent information (Dempsey, Driscoll, & Swindell, 1993). Nevertheless, there are certain circumstances that seem to favour delayed feedback, such as where it is associated with a later event of reteaching, or extending the teaching of the language component that prompted the feedback, whereby more benefit emerges from reteaching through delayed feedback.

Immediate feedback is one of the strengths of AI, compared with a human teacher who may take several weeks to review the essays produced by an entire class and return them with feedback (Steiss et al., 2024).

*2.2 Use of AI for Writing Feedback*

AI has existed is some form, with claims to provide feedback on (and/or the scoring of) writing, since the 1960s. It is termed *automated writing evaluation*, which is "designed to provide instant computer-generated scores for a submitted essay along with diagnostic feedback" (Chen & Cheng, 2008, p. 94).

An early attempt was Page's (1966) Project Essay Grade, which assigned essay scores based on predefined criteria. Then, in 1997, Foltz and Landauer produced the Intelligent Essay Assessor to assess undergraduate essays (Foltz, Laham, & Landauer, 1999). Until recently, however, these programs were mostly designed to handle English writing by native speakers, and relied primarily on simple statistical measures of word and sentence length, and spelling errors, as well as approximate indicators of content and organisation. Although automated, they were not particularly *intelligent*. Nevertheless, they led to the emergence of various commercial applications and platforms (e.g., Grammarly, Turnitin, and WriteLab), as well as essay grading systems (e.g., E-rater, Criterion, and MI Write) (Do, Kim, &

Lee, 2023), which are now beginning to incorporate modern AI (e.g., ChatGPT) and could more credibly replace a human scorer and feedback provider in a foreign language teaching setting.

Research (e.g., Nazari, Shabbir, & Setiawan, 2021) testifies to the increased objective accuracy of such scoring over the years, as well as the increased ability of software to provide useful feedback. However, even the latest generative AI is recorded as evidencing certain aspects that would negatively impact its ability as a provider of scores and feedback. For example, it does not always recognise or apply logical reasoning correctly, and can generate and accept incorrect information (Zhou, Qiu, Huang, & Zhang, 2023), which is sometimes referred to as *hallucination*. In addition, the quality of generative AI can be quite sensitive to small differences in the wording of the human prompt provided, in terms of specifying what is required (Steiss et al., 2024). Such findings underscore the need for studies that explore in greater depth the use of generative AI for a specific pedagogical purpose. At the time of data collection, it was believed that the present study had no close parallel. However, Steiss et al. (2024) have since published their research with a very similar theme but in a different context, which will thus serve as a useful reference for comparison in this study.

## 3. Research Questions

Based on the above review, and taking into account more general considerations also covered by Steiss et al. (2024), the following research questions were established with respect to ChatGPT:

Q1. Is the feedback consistent? (including being underpinned by a systematic set of criteria)

Q2. Is the feedback credible? (i.e., in some sense accurate)

Q3. What types of feedback are provided, in terms of the dimensions reviewed above?

## 4. Method

Document analysis was utilised, which is a systematic approach to qualitative research that is focused on written text (Bowen, 2009). Furthermore, in terms of its reliance on expert judgment, the study has parallels with Steiss et al. (2024).

### 4.1 Participants

In total, 29 English major students at a provincial university in the Kingdom of Saudi Arabia (KSA) agreed for their essays to be employed in the research. They were in their second year of undergraduate study, and the essays were produced for the Writing II module. In order to ensure confidentiality, all essays and feedback are referred to by a code (e.g., E10, F10).

### 4.2 Elicitation of Essays

The essays were produced in a regular writing class as a form of writing practice. The writing prompt was to produce a five-paragraph cause-and-effect essay, following the taught approach, about a topic from their textbook: *The positive and negative health effects of exercising*. After an initial teacher-led brainstorming of the topic, the students were free to write without time pressure, but with the absence of any additional sources of assistance, for the remainder of the class period. Although a convenience sample, the essays may be regarded as typical of those produced by students at this level in the KSA in terms of their language and content.

### 4.3 Elicitation of Feedback

In the present study, there was a specific issue that rendered this stage more challenging than the norm in most contemporary writing classes, that is, the teacher in this class habitually elicited writing in handwritten form, rather than word-processed. This was partly due to an ongoing tradition at the university (although not representative of the practice at all universities in the KSA), and partly because it removed some unwanted issues that may accompany word-processing, such as the students using spell- or grammar-checking software, and accessing the internet for texts to replicate in their work. Nevertheless, this meant that the student essays were not initially in the electronic form required by generative AI applications such as ChatGPT.

Experimentation revealed that optical character recognition software was not suitable for the conversion of the handwritten scripts into digital text for the purposes of this study, since it either failed to decode the poor quality handwritten English, or did so with a degree of tidying up and correction that was obviously undesirable in the study context. Therefore, since the essays were short (average length: 172 words), the researcher manually created word-processed versions, while carefully retaining all of the errors including, for example, many failures to observe spacing between words or the distinction between upper and lower case.

Following that, the electronic versions were submitted, in turn, to ChatGPT with the following prompt: "*A Saudi learner of English as a foreign language was asked to write a five-paragraph cause-effect essay about the positive and negative health effects of exercising. Please give some helpful feedback in simple English. Here is what was written:*".

This prompt was chosen in order to provide key information about the students' writing while not suggesting anything highly specific about the type of feedback required. Thus, there was no mention of the correction of errors or focus on content, for example.

### 4.4 Data Analysis

The researcher conducted qualitative analysis of both the essays and the feedback. The replicability of the analysis was assessed by the researcher re-analysing the data after a break of two months, in order to discover any differences (intra-coder reliability). Content validation of the analysis was achieved by cross-checking with an independent native speaker and expert in the field. The essays were analysed for key features such as the number of paragraphs and function of each (organisation), the types of language error, and the coverage of the targeted

content, which are typical features of focus in EFL compositions at this level (Tiwari, 2023). The feedback was analysed for consistency and credibility (Lincoln & Guba, 1985), and for the presence of each of the features described in the literature review. This represents a considerably more granular approach (albeit at a smaller scale in terms of the number of participants) than Steiss et al. (2024), who also conducted expert analysis of ChatGPT feedback on students' writing.

## 5. Results

The ChatGPT feedback was impressive in terms of its faultless English and, on overall impression, the detailed and useful feedback that was simply presented. However, deeper analysis revealed a more nuanced picture.

### 5.1 Consistency

ChatGPT is consistent in that, given the same prompt, it closely follows the same pattern of response for each essay. It presented its feedback in a series of labelled, self-selected, categories – Introduction, Organization, Spelling and Grammar, Clarity, Content, and Conclusion – that were nearly always reported in the same order, but usually not all were used, with the Conclusion and Introduction the least included categories. Furthermore, the feedback provided in each category was mostly echoed between texts.

The least successful feature of the chosen categories was that the Clarity and Content categories tended to overlap, that is, Clarity often included suggestions about the need for more specific information, while for other texts these suggestions were mentioned under the Content category. The worst instance of this was F10, where these two categories essentially reported identical feedback:

1. Clarity:

   - The essay seems to discuss the importance of mental health and the role of both medical and religious approaches but doesn't delve into the cause-and-effect relationship between exercising and mental health.

2. Content:

   - While the essay discusses the importance of mental health and the coexistence of medical and religious approaches, it doesn't address the topic of exercise and its impact on mental health, which is the main theme of the essay.

Moreover, the feedback is not consistent in that, if asked to comment twice on the same essay, and with the same prompt, the responses do not match exactly. However, the differences echo those which a human teacher might exhibit if asked to provide feedback on the same essay after an interval of time, whereby they may not comprehensively recall and reproduce their earlier feedback (intra-judge reliability).

### 5.2 Credibility

ChatGPT provides predominantly credible feedback in response to the essay content, and the specified genre and organisation. This failed only in a few instances, which all involved higher-level rhetorical structure rather than detailed language points. For example, in the case of F3:

1. Organization:

   - The essay should have a clear structure with five paragraphs: introduction, positive health effects of exercising, negative health effects, the importance of moderation, and a conclusion. This will make your ideas more organized.

While fine in isolation, especially since ChatGPT was not informed of the exact focus required for each paragraph, the feedback does not entirely suit E3 which does have five paragraphs, and (although the English spelling and grammar are poor) the information in each paragraph does fit the requirements, with the exception that paragraph 4 repeats the negative aspects of paragraph 3 rather than encouraging moderation.

Again, in the case of F2:

1. Introduction:

   - The introduction is good, but you could make it more engaging by starting with a compelling fact or question about exercise.

In fact, although poorly worded, the E2 introduction is engaging by presenting a paradox, and indeed poses a question:

> *Did you know that exercising could be unheathy? There are several benefits of exercising and sometimes it could be harmful.*

Then, again since E26 in fact has two paragraphs arguing in favour, two arguing against, and then one concluding the argument in favour of moderation, the following comment in F26 is slightly off target:

1. Organization:

   - The essay could be better organized by separating the positive and negative effects of exercise into distinct paragraphs.

Here, ChatGPT is evidencing the fault seen more widely, whereby it is capable of claiming a highly plausible issue, but one that does not actually exist in the essay.

*5.3 Corrective Feedback Response*

CF was absolutely the most widely used feedback approach detected, with similar attention given across language, organisation, and content. The tendency was to supply a full correction for the language points addressed, and strong suggestions about the content and organisation points highlighted, in order to provide the writer with clarity about the correction. Thus, self-correction is only moderately invoked. To the reader of the feedback, this *voice* locates ChatGPT clearly in the role of a teacher or writing expert, acting independently.

The content and organisation feedback is often of this type, providing strong hints but not rewriting the text (which ChatGPT can, of course, also carry out if requested explicitly in the prompt). For example, in the case of F2:

1. Clarity:
   - In the second paragraph, explain more about how exercise improves brain health, prevents cancer, and makes you more active. Provide specific examples or reasons.
   - When discussing the negative effects of too much exercise, be specific about how it makes muscles sore and affects the immune system.
   - In the fourth paragraph, clarify what a "healthy exercising schedule" means and provide tips for creating one.

It is only when unable to decode the meaning that ChatGPT can make no suggestion, as seen here in the case of F8:

1. Clarity:
   - In the first paragraph, clarify the statement about "70% of people facing death." It's not clear what this means in the context of the essay.

The original E8 text was *Do you Know that 70% of PeoPle face death every day because of exercises!*, which even to a human reader is decidedly unclear due to the unexpectedness of the pragmatics.

Interestingly, with respect to low-level correction, when ChatGPT cites the words of the essay it usually does so exactly, rather than correcting any of the errors, as seen in E1:

1. Content:
   - In the introduction, you mention "bieng active Person," but you don't explain what that means.

An opportunity is thus missed to embed in the content correction a correction of spelling/grammar via recasting, by writing: *being an active person*.

In other respects, the low level correction is all of the complete type that supplies the word(s) requiring attention and the full correction, thus negating the requirement for the writer to carry out self-correction, for example (F19):

1. Spelling and Grammar:

   Correct "kePt" to "kept," "vulnarable" to "vulnerable," "feending" to "depending,"

*5.4 Communicative Response*

As stated above, ChatGPT remains firmly in the role of a teacher or expert commentator, rather than a speaker in a non-pedagogical message-oriented conversation. The discourse is conversation-like in the use of contracted forms like *doesn't* (n=26) and some use of second-person address (e.g., explicit *you*, or *you* implied in an imperative). However, often the more detached third-person (e.g., *The essay* or *it*) is used, as seen in this response to the E16 writing:

1. Content:
   - The essay emphasizes the importance of daily exercise and its benefits but doesn't delve into the negative effects of not exercising.

Even where *you* is implied, the message is still not that of a conversational interlocutor, for example, in the case of E3:

1. Content:
   - In the fourth paragraph, clarify why too much exercise is not right for health.

For example, the more fully communicative response might be: *I would love to know why too much exercise can be bad for your health*.

A final aspect in which the feedback is not communicative is the tendency in places, especially for Content, to provide a response that simply contains a paraphrase or summary of the student's writing (the kind of response referred to by Steiss et al. (2024) as *fluff*). Such feedback is neither particularly communicative nor useful to the writer (although it might be helpful for a human teacher reading the ChatGPT feedback), because it conveys no new information. For example, consider this response to the F27 essay:

1. Content:
   - The essay discusses the significance of health and why people ask about your health when they meet you.
   - It talks about how people often neglect their health due to various factors like busy work schedules, poor diets, and climate changes.

- The essay touches on the importance of staying healthy to prevent diseases like heart disease and cancer.

There are some contexts that would provide such repetition with a communicative function, but these would require a different framing, for example, if the speaker wanted to register that they agree with the writer (a highly communicative response); however, they would need to cast this more as *I quite agree with you about the significance of health ... and ...* . Another function could be the pedagogical one termed *recast*, where the teacher repeats back to the student the correct form of the production, thus engaging in covert correction. Indeed, the E27 essay was characterised by numerous low-level errors. The second bullet point of feedback above corresponds to: *mostly people became ill because of burdenization of work,eating unhealthy Food change inclimate. etc.* However, if this feedback in F27 is to be understood as a recast, it needs to be introduced rather by *You mean ... ?*, rather than *It talks about how … .*

*5.5 Emotional/Affective Response*

The extent of the emotional/affective response is limited. Generally, the affective dimension is represented covertly by the overtone from the correctness feedback on the language or content and organisation. For example, the language feedback was primarily focused on the correction of errors as opposed to the recognition of effective points, and thus affectively negative. The content and organisation feedback however in many cases recorded a mixture of good points and errors, and therefore was partially positive in terms of the emotional effect.

Only rarely was a component of the essay explicitly stated as good, for example in the case of F5:

1. Introduction:

   - The introduction is a good start, but you can improve it by being more explicit about the topic. Begin by introducing the concept of positive and negative health effects of exercising.

Far more frequently, an effective aspect was only implied as being good, as in this part of the Clarity feedback F13, which appears to simply repeat part of the essay content, but in fact is implying correctness and thus is emotionally positive:

1. Clarity:

   - The essay provides a basic overview of the causes and effects of exercising on health. However, it could benefit from more specific examples and details.

In these, as per many examples in the feedback, a mention of something good (at least implicitly) precedes a negative feature, often linked with a discourse marker like *but* or *however*. Moreover, this represents a common strategy of human feedback providers through the aforementioned feedback sandwich, where the emotive effect of a negative point is softened by preceding and subsequent positive observations.

*5.6 Generality*

The feedback tended to generalise about the entire essay much more frequently with respect to content than lower-level language matters, although unusually in the case of F29 both facets were addressed:

1. Spelling and Grammar:

   - The essay has some minor grammar issues and misspellings.

2. Clarity:

   - The essay is clear in presenting the positive and negative aspects of exercising.

*5.7 Completeness*

Similar to many human feedback providers, ChatGPT did not provide complete feedback, which was most noticeable at the low level where many errors of spelling, capitalisation, punctuation, vocabulary, and grammar were often omitted. In the E3 essay, for example the researcher identified 44 errors in terms of spelling (n=14), punctuation (n=3), capitalisation (n=5), grammar (n=18), and vocabulary (wrong word, n=4), whereas the feedback only addressed five specific issues:

1. Spelling and Grammar:

   - Use "maintained" instead of "maintaided."
   - Correct "sPort" to "Sport."
   - Fix "daialy" to "daily."
   - Use "a happy life" instead of "haPPy life."
   - Add periods at the end of sentences for proper punctuation.

Although incomplete, this strategy may represent a sensible choice in many instances, although the reason for selecting those five points is unclear. For instance, *exercise* was used five times and spelt incorrectly each time, but was not included in the feedback, while a period that was missing at the end of only one sentence was included as feedback. Grammar (in a broad sense) was not mentioned despite several basic errors such as *they will be have some problem*, *eat a healthy food*, *not right for them health*, and *the sport help us.* In fact, this E3 writer is remarkable for quite sophisticated vocabulary choice (i.e., *maintain*, *mood*, *health*, and *problem*), but with the spelling and

grammar of a lower proficiency writer. Thus, it is difficult to regard ChatGPT's selection of what to give feedback on here as anything other than random.

## 6. Discussion

The present study complements the other major investigation in the same research area (Steiss at al., 2024) by demonstrating in greater qualitive detail the manner in which ChatGPT differs from, and in some cases falls slightly short of a human feedback provider. The consistency (across texts) and credibility (objective truth) of the feedback were shown to be good. Otherwise, ChatGPT predominantly adopted the role of a teacher, and the feedback was found to be primarily corrective, focused in a balanced manner on the content and organisation, as well as lower-level language points. Affective feedback was not separately highlighted and truly communicative feedback was almost entirely absent. The feedback was targeted at providing correct forms for lower-level language features, and offering strong suggestions for issues at the content and organisation levels. Full completeness of feedback was not delivered, which is not implied as necessarily representing a deficiency. In those respects, the behaviour of many teachers as feedback providers was echoed. However, the choice of what lower-level items to provide feedback on did not match that which a dedicated teacher might select. The broad finding of Steiss et al. (2024) is therefore supported, in that ChatGPT falls short of the feedback provided by an expert human teacher.

There is more detailed similarity between the results of this study and those of Steiss et al. (2024), whereby they also found the greatest disparity between human and ChatGPT feedback with respect to accuracy and the prioritisation of essential features (especially selecting the most important errors to comment on). Moreover, this similarity is perhaps notable, given that Steiss at al. (2024) targeted a quite different writer context and population (i.e., school children in California), meaning that the essays elicited were somewhat different.

The main implication for teachers of writing is that feedback such as that elicited from ChatGPT does closely emulates what many teachers provide, which is likely due to the ChatGPT training material having contained examples of feedback from such teachers. Therefore, if teachers seek feedback with different emphases, they will need to create appropriate elicitation prompts. Essentially, they will need to become, albeit in a limited fashion, what is referred to as *prompt engineers*, namely, experts in crafting requests addressed to AI software that they yield the precise information required.

However, problems remain in terms of the occasional falsity of the feedback, or random rather than motivated choices of what to include, which teachers will need to review themselves, accept, or address through their development as prompt engineers. Moreover, as Steiss et al. (2024) note, there is ultimately a trade-off for the teacher between the slightly inferior quality of the feedback and the time benefits derived from its generation. The latter are twofold. The teacher saves in time reading and providing the feedback, and the student benefits from the fact that the feedback can be received far sooner than from the typical teacher.

The fact that feedback of this quality can even be made available to a writer without a teacher, at the very moment after writing, just as spellchecking is, would usually be considered a benefit (Dempsey et al., 1993). However, it remains to be researched whether that would generate learning more effectively than feedback slightly delayed and mediated through a teacher, as simulated in the present study. It must not be forgotten that the educational goal of feedback is usually long-term learning rather than just improvement of the current essay draft (see 2.1.1). There is a risk that instant feedback, maybe even without a record kept for the teacher of what it was, does not generate enough depth of processing (Craik, 2002) to promote the long-term retention that constitutes genuine language learning.

## 7. Conclusion

This study is of course limited to one AI tool, accessed at one particular point in time using just one generalised prompt. Therefore, it is pertinent that using other prompts specifying, for example, particular aspects of the essay to focus on, would likely have significantly altered the feedback provided. Furthermore, ChatGPT is under constant development, and thus a teacher or researcher accessing the tool today will not necessarily obtain the same kind of feedback as reported here. Nevertheless, the obtained responses to the research questions are suggestive of the general capabilities of AI at the present time, and of which teachers require awareness. This remains the case whether one employs AI more directly (e.g., by using ChatGPT as per the present study) or indirectly through software that itself accesses resources such as ChatGPT in support of its own provision (e.g., Grammarly and Cograder).

### Authors' contributions

Dr. Manal Alghannam was the only one who was responsible for the research design, data collection, discussion, and manuscript writing.

### Competing interests

I declare that I did not use any personal relationship in my study to influence the results of the conducted research.

### Informed consent

Not applicable

**Ethics approval**

The Publication Ethics Committee of the Sciedu Press.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

**Provenance and peer review**

Not commissioned; externally double-blind peer reviewed.

**Data availability statement**

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

**Data sharing statement**

No additional data are available.

**Open access**

**Copyrights**

**References**

Andersen, P. A., & Andersen, J. F. (2005). Measurements of perceived non-verbal immediacy. In V. Manusov (Ed.), *Sourcebook of non-verbal measures: Going beyond words* (pp. 113-126). Routledge.

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment and Evaluation in Higher Education*, *38*(6), 698-712. https://doi.org/10.1080/02602938.2012.691462

Bowen, G. A. (2009). Document analysis as a qualitive research method. *Qualitative Research Journal*, *9*(2), 27-40. https://doi.org/10.3316/QRJ0902027

Carless, D. (2020). From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes. *Active Learning in Higher Education*, *23*(2), 143-153. https://doi.org/10.1177/1469787420945845

ChatGPT. (2023). Retrieved from: https://chatgpt.com

Chen, C. F., & Cheng, W. Y. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, *12*(2), 94-112.

Collin, J., & Quigley, A. (2021). *Teacher feedback to improve pupil learning*. London, UK: Education Endowment Foundation.

Cotos, E. (2023). Automated feedback on writing. In O. Kruse, C. Rapp, C. M. Anson, K. Benetos, E. Cotos, A. Devitt & A. Shibani (Eds.), *Digital writing technologies in higher education*. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-031-36033-6_22

Craik, F. (2002). Levels of processing: Past, present and future? *Memory, 10*(5/6), 305-318. https://doi.org/10.1080/09658210244000135

Dempsey, G. V., Driscoll. M. P., & Swindell, L. K. (1993). Text based feedback. In G. V. Dempsey & G. C. Sales (Eds.), *Interactive instruction and feedback* (pp. 21–54). Englewood Cliffs, NJ: Educational Technology.

Devine, A. (2018). *Language tests for children – gentle encouragement or too much stress?* Cambridge University Press. Retrieved from https://www.cambridgeenglish.org/blog/benefits-of-cambridge-assessment-english-tests-for-young-learners/

Diab, N. M. (2016). A comparison of peer, teacher and self-feedback on the reduction of the language errors in student essays. *System, 57,* 55-65. https://doi.org/10.1016/j.system.2015.12.014

Didau, D., & Rose, N. (2016). *What every teacher needs to know about psychology*. John Catt Publisher.

Do, H. J., Kim, Y. S., & Lee, G. G. (2023). *Prompt- and trait relation-aware cross-prompt essay trait scoring*. https://doi.org/10.18653/v1/2023.findings-acl.98

Edge, J. (1989). *Mistakes and correction*. London, UK: Longman.

Education Endowment Foundation. (2022). *Feedback*. Retrieved from https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit/feedback

El Nokali, N. E., Bachman, H. J., & Votruba-Drzal, E. (2010). Parent involvement and children's academic and social development in elementary school. *Child Development*, *81*(3), 988-1005. https://doi.org/10.1111/j.1467-8624.2010.01447.x

Foltz, P., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer - Enhanced Learning*, *1*(2).

Han, Z. H. (2002). Rethinking the role of corrective feedback in communicative language teaching. *RELC Journal*, *33*(1). https://doi.org/10.1177/003368820203300101

Hanif, H. (2020). *The role of L1 in an EFL classroom*. The Language Scholar. University of Leeds. Retrieved from https://languagescholar.leeds.ac.uk/the-role-of-l1-in-an-efl-classroom/

Harlen, W., & Deakin Crick, R. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI-Centre Review, version 1.1*). *Research Evidence in Education Library* (Vol. 1). London, UK: EPPI-Centre, Social Science Research Unit, Institute of Education.

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

Kaye, P. (2024). *Making writing communicative*. British Council. Retrieved from https://www.teachingenglish.org.uk/professional-development/teachers/understanding-learners/articles/making-writing-communicative

Lightbown, P. M., & Spada, N. (2006). Explaining second language learning. In P. M. Lightbown & N. Spada (Eds.), *How languages are learned* (pp. 29-50). Oxford, UK: Oxford University Press.

Lin, H. C. K., Wu, C. H., & Hsueh, Y. P. (2014). The influence of using affective tutoring system in accounting remedial instruction on learning performance and usability. *Computers in Human Behavior*, *41*, 514-522. https://doi.org/10.1016/j.chb.2014.09.052

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage Publications. https://doi.org/10.1016/0147-1767(85)90062-8

Lipnevich, A., & Panadero, E. (2021). A review of feedback models and theories: Descriptions, definitions, and conclusions. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.720195

Lundgren, L. (2022). *The correctness of corrective feedback*. Örebro, Sweden: Örebro University, Department of Humanities, Education and Social Sciences.

Maolida, E. H. (2017). Relating teacher's oral corrective feedback to young learners' uptake: A case study in a young learner EFL classroom. *Indonesian EFL Journal*, *3*(2), 181-192. https://doi.org/10.25134/ieflj.v3i2.665

Mayordomo, R. M., Espasa, A., & Guasch, T. (2022). Perception of online feedback and its impact on cognitive and emotional engagement with feedback. *Educational Information Technology*, *27*, 7947-7971. https://doi.org/10.1007/s10639-022-10948-2

Meunier, F., & Muñoz, F. S. (2022). Teacher feedback in the foreign language classroom: Navigating between research findings, beliefs, and classroom practices. In *Part III: Applied linguistics & foreign language pedagogy* (pp. 271-295). Brussels, Belgium: Presses de l'Université Saint-Louis. https://doi.org/10.4000/books.pusl.27976

Molloy, E. (2010). The feedforward mechanism: A way forward in clinical learning? *Medical Education*, *44*, 1157-1159. https://doi.org/10.4000/books.pusl.27976

Mory, E. (2001). Feedback research. In *The handbook of research for educational communications and technology* (pp. 919-956). Bloomington, Indiana: Association for Educational Communications and Technology.

Nagata, N. (1993). Intelligent computer feedback for second language instruction. *The Modern Language Journal*, *77*(3), 330-339. https://doi.org/10.1111/j.1540-4781.1993.tb01980.x

Nazari, N., Shabbir, M. S., & Setiawan, R. (2021). Application of artificial intelligence powered digital writing assistant in higher education: Randomized controlled trial. *Heliyon*, *7*(5), e07014. https://doi.org/10.1016/j.heliyon.2021.e07014

Oxford Languages. (2022). *Feedback*. Oxford University Press. Retrieved from https://languages.oup.com/google-dictionary-en/

Page, E. B. (1966). The imminence of ... grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238-243.

Panova, I., & Lyster, R. (2002). Patterns of corrective feedback and uptake in an adult ESL classroom. *TESOL Quarterly*, *36*(4). https://doi.org/10.2307/3588241

Paulmann, S., & Weinstein, N. (2023). Teachers' motivational prosody: A pre-registered experimental test of children's reactions to tone of voice used by teachers. *Educational Psychology*, *93*(2), 437–452. https://doi.org/10.1111/bjep.12567

Reich, R. (2022, 28 Nov). Now AI can write students' essays for them, will everyone become a cheat? *The Guardian*. Retrieved from: https://www.theguardian.com/commentisfree/2022/nov/28/ai-students-essays-cheat-teachers-plagiarism-tech

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and methods in language teaching*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511667305

Russell, V. (2009). Corrective feedback, over a decade of research since Lyster and Ranta (1997): Where do we stand today? *Electronic Journal of Foreign Language Teaching*, *6*(1), 21-31.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119-144.

https://doi.org/10.1007/BF00117714

Sales, G. C. (1993). Adapted and adaptive feedback in technology-based instruction. In G. V. Dempsey & G. C. Sales (Eds.), *Interactive instruction and feedback* (pp. 159-175). Englewood Cliffs, NJ: Educational Technology.

Schachter, J. (1991). Corrective feedback in historical perspective. *Second Language Research*, *7*, 89-102. https://doi.org/10.1177/026765839100700202

Schweers, C. W. (1999). Using L1 in the L2 classroom. *English Teaching Forum*, *37*, 6–13.

Scott, S. V. (2014). Practising what we preach: Towards a student-centred definition of feedback. *Teaching in Higher Education*, *19*(1), 49-57. https://doi.org/10.1080/13562517.2013.827639

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., … Warschauer, M. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, https://doi.org/10.1016/j.learninstruc.2024.101894

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, *16*(3), 371-391. https://doi.org/10.1093/applin/16.3.371

Taylor, R. P. (1980). *The computer in the school: Tutor, tool, tutee.* New York, NY: Teachers' College Press.

Terzis, V., Moridis, C., & Economides, A. (2012). The effect of emotional feedback on behavioral intention to use computer based assessment. *Computers and Education*, *59*(2), 710-721. https://doi.org/10.1016/j.compedu.2012.03.003

Tiwari, H. P. (2023). Use of Jacobs ESL Composition Profile to evaluate university students' writing. *ELT Worldwide Journal of English Language Teaching*, *10*(2), 288. https://doi.org/10.26858/eltww.v10i2.51632

Uysal, H. H. (2010). The role of grammar and error correction in teaching languages to young learners. In B. Haznedar & H. H. Uysal (Eds.), *Handbook for teaching foreign languages to young learners in primary schools* (pp. 233-252). Ankara, Turkey: Anı Publications.

Vigil, N. A., & Oller, J. W. (1979). Rule fossilization: A tentative model. *Language Learning*, *26*, 281-295. https://doi.org/10.1111/j.1467-1770.1976.tb00278.x

Ware, P. D., & O'Dowd, R. (2008). Peer feedback on language form in telecollaboration. *Language Learning and Technology*, *12*(1), 43-63.

Wirantaka, A. (2019). Investigating written feedback on students' academic writing. *Advances in Social Science, Education and Humanities Research*, *353*, 1-7. https://doi.org/10.2991/icosihess-19.2019.1

York, B. N. (2014). *Know the child: The importance of teacher knowledge of individual students' skills (KISS).* Stanford, CA: CEPA Working Paper, Center for Education Policy Analysis at Stanford University.

Zhou, J., Ke, P., Qiu, X. P., Huang, M. L., & Zhang, J. P. (2023). ChatGPT: Potential, prospects, and limitations. *Frontiers of Information Technology & Electronic Engineering*, *25*, 6-11. https://doi.org/10.1631/FITEE.2300089