CADERNOS de LINGUAGEM & SOCIEDADE

Papers on Language and Society

ChatGPT-supported formative assessment in interpreter education: validity of automated ratings and perceived quality of diagnostic feedback

Avaliação formativa apoiada pelo ChatGPT no ensino de interpretação: validade das classificações automáticas e qualidade percebida do feedback diagnóstico

Evaluación formativa apoyada por ChatGPT en la enseñanza de la interpretación: validez de las calificaciones automáticas y calidad percibida de la retroalimentación diagnóstica

Wenjing Liu
13986435897@163.com
https://orcid.org/0009-0004-7865-5549
Macao Polytechnic University, China

Adriana Silvina Pagano
apagano@letras.ufmg.br
https://orcid.org/0000-0002-3150-3503
Universidade Federal de Minas Gerais (UFMG), Minas Gerais, Brasil

ABSTRACT

Formative assessment plays a critical role in teaching and learning. Recent advances in large language models (LLMs) have enabled their application as automated assessment systems and feedback providers. This study explores the validity of ChatGPT-based assessment and the perceived quality of its feedback in interpreter education. To this end, ChatGPT-4o was used to assess 60 Chinese–Portuguese simultaneous interpreting tasks, producing rubric-based quantitative ratings and qualitative diagnostic feedback. To examine its effectiveness, three types of validity (concurrent, predictive, and know-group) were examined by comparing ChatGPT-generated scores with those of nine trained human raters. A *post-hoc* questionnaire was also administered to collect raters' subjective perceptions of the feedback. Results show strong alignment between the model and human scores, with ChatGPT demonstrating robust predictive power and discriminative ability. Raters viewed the feedback favorably and supported its use as a complement to teacher feedback, highlighting the pedagogical value of LLMs in interpreter training.

Keywords: formative assessment; automated scoring; feedback; large language model; ChatGPT; interpreter education.

Recebido em: 25 de setembro de 2025
Aceito em: 24 de fevereiro de 2026
DOI: 10.26512/les.v26i2.59793

ARTIGO

RESUMO

A avaliação formativa desempenha um papel fundamental no ensino e na aprendizagem. Os avanços recentes nos modelos de linguagem de grande escala (LLMs) permitem a sua aplicação como ferramentas para a avaliação automática e o fornecimento de feedback. Este estudo explora a validade da avaliação baseada no ChatGPT e a qualidade percebida do seu feedback no contexto da formação de intérpretes. Para esse propósito, o ChatGPT-4o foi utilizado para avaliar 60 tarefas de interpretação simultânea chinês-português, produzindo classificações quantitativas baseadas numa rubrica e feedback qualitativo. Para avaliar a sua eficácia, foram examinados três tipos de validade (concorrente, preditiva e de grupos conhecidos), comparando-se os resultados gerados pelo ChatGPT com os de nove avaliadores humanos treinados. Um questionário pós-tarefa também foi aplicado para recolher perceções subjetivas dos avaliadores sobre o feedback. Os resultados revelam uma forte concordância entre as pontuações do modelo e as humanas. O ChatGPT demonstra um elevado poder preditivo e capacidade discriminativa. Os avaliadores avaliaram positivamente o feedback e apoiaram a sua utilização como complemento ao feedback dos professores, sublinhando o valor pedagógico dos LLMs na formação de intérpretes.

Palavras-chave: avaliação formativa; classificação automática; feedback; modelos de linguagem de grande escala; ChatGPT; ensino de interpretação.

RESUMEN

La evaluación formativa desempeña un papel fundamental en la enseñanza y el aprendizaje. Los avances recientes en los modelos de lenguaje de gran escala (LLM) permiten su aplicación como herramientas para la evaluación automática y la provisión de retroalimentación. Este estudio explora la validez de la evaluación basada en ChatGPT y la calidad percibida de su retroalimentación en el contexto de la formación de intérpretes. Con este propósito, se utilizó ChatGPT-4o para evaluar 60 tareas de interpretación simultánea chino-portugués, generando puntuaciones cuantitativas basadas en una rúbrica y retroalimentación cualitativa. Para valorar su eficacia, se examinaron tres tipos de validez (concurrente, predictiva y de grupos conocidos), comparando los resultados generados por ChatGPT con los de nueve evaluadores humanos formados. También se aplicó un cuestionario posterior a la tarea para recoger las percepciones subjetivas de los evaluadores sobre la retroalimentación. Los resultados revelan una fuerte concordancia entre las puntuaciones del modelo y las humanas. ChatGPT muestra un alto poder predictivo y una sólida capacidad discriminativa. Los evaluadores valoraron positivamente la retroalimentación y respaldaron su uso como complemento a la retroalimentación del profesorado, subrayando el valor pedagógico de los LLM en la formación de intérpretes.

Palabras clave: evaluación formativa; calificación automática; retroalimentación; modelos de lenguaje de gran escala; ChatGPT; enseñanza de interpretación.

Como citar:

LIU, Wenjing; PAGANO, Adriana Silvana. ChatGPT-supported formative assessment in interpreter education: validity of automated ratings and perceived quality of diagnostic feedback. **Cadernos de Linguagem e Sociedade**, Brasília, v. 26, n. 2, p. 265-290, jul./dez. 2025. DOI: 10.26512/les.v26i2.59793 Disponível em: . Acesso em: XXX.

Correspondência:

Nome por extenso do autor principal

Rua XXX, número XXX, Bairro XXX, Cidade, Estado, País.

INTRODUCTION

Assessment plays a central role in interpreter education, not only as a means of measuring learner performance but also as a tool to guide learning and teaching (Han, 2022; Sawyer, 2004; Scaramucci, 2006). Among various types of assessment, formative assessment is particularly valued for its pedagogical utility: it offers learners ongoing opportunities to monitor their learning process (Boud; Soler, 2016) and allows teachers to make informed adjustments during the teaching process (Alahmadi *et al*., 2019).

According to Wiliam and Thompson (2017), formative assessment can be enacted through several key strategies, such as clarifying success criteria, promoting classroom discussions, and providing feedback. In this regard, as a fundamental mechanism of formative assessment, feedback serves as a pedagogical resource that facilitates learner reflection, promotes skill development, and enhances motivation (Crezee; Grant, 2016; Yan; Amini; Kasuma, 2023). However, assessing students' performance and, accordingly, producing high-quality formative feedback is labor-intensive and time-consuming (Han; Lu; Fan, 2025). While instructors often struggle to deliver timely feedback, students express a strong need for detailed responses, not only for academic improvement but also for emotional support (Lee, 2018). Furthermore, learners have different preferences and expectations of feedback (Tahraoui, 2022), making the provision of regular feedback a persistent challenge in interpreter education.

Recent advances in large language models (LLMs) such as GPT have opened new possibilities for automating the assessment and feedback provision. In areas such as second-language writing and translation, studies have demonstrated the feasibility of using LLMs for such task (Imran; Almusharraf, 2023; Teng, 2024; Xu; Su; Liu, 2025). Yet, interpreter education poses unique challenges. Unlike text-based writing or translation, interpreting involves spoken input, making automated evaluation more technically demanding, particularly when assessing fluency. Nevertheless, recent work has begun to explore text-based approaches to interpreting assessment, using transcripts of source and target speeches as input for LLMs to approximate interpreting assessment (Han; Lu; Fan, 2025; Jia; Aryadoust, 2024; Wang; Wang, 2025). These studies demonstrate the potential of LLMs in producing relevant assessment even without access to interpretation audios.

Despite these promising developments, the current literature remains limited in scope. Most existing research focus on consecutive interpreting and involves a narrow range of language pairs, and very few studies have systematically examined the content and quality of the feedback generated by LLMs, leaving its educational consequences largely unexplored. To address these gaps, the present study investigates the use of ChatGPT as an automated assessment system and feedback provider in Chinese-to-Portuguese simultaneous interpreting training. Our approach integrates both quantitative scores and qualitative evaluative comments, aiming to support learners

in gauging their performance levels, understanding the rationale behind the assessment, and identifying areas for improvement. Within this framework, we examine the validity of the scores generated by ChatGPT-4o and explore how human raters perceive the quality and the potential impacts of its feedback. In doing so, this study provides empirical evidence to inform the pedagogical use of LLM-based tools in interpreter education.

1. LITERATURE REVIEW

This section outlines two major strands of related work: formative assessment feedback in the context of interpreter training, and recent advances in the automatic assessment of interpreting quality. These strands help contextualize the role of feedback in interpreting education and highlight the potential of technology-assisted assessment methods.

1.1 Formative assessment feedback in interpreter training

In educational contexts, assessment can take different forms depending on its purpose. Summative assessment evaluates learning outcomes at the end of an instructional period and is typically used for grading or certification, with limited impact on the learning process itself (Glazer, 2014). Formative assessment, by contrast, is integrated into ongoing instruction, bridging the gap between assessment, and learning by tracking individual student progress and fostering learners' self-regulatory capacities (Boud; Soler, 2016). Within this framework, assessment is conceptualized as a mechanism that enhances the quality of learning (Damacena; Quevedo-Camargo, 2021), boosts learner motivation (Yan *et al*., 2023), and helps instructors to adapt teaching methods and content to better meet learners' needs (Alahmadi *et al*., 2019).

A central component of formative assessment is feedback. As Kelly (2014) notes, formative assessment consists of both evaluative and corrective comments on student performance, serving as detailed feedback to support learning or task performance. Feedback is broadly defined as information provided by an agent (e.g., teacher, peer, textbook, parent, learner) regarding aspects of a learner's performance (Hattie; Timperley, 2007), which underscores the diversity of potential feedback sources. In the context of interpreting pedagogy, feedback can be produced by teachers, peers, or the students themselves. Among these sources, teacher feedback is generally preferred by students (Lee, 2018; Xue; Liu, 2024), whereas peer and self-feedback show different evaluative biases across studies (Fowler, 2007; Holewik, 2020; Wu, 2017), highlighting the need to enhance students' assessment literacy. In terms of content, interpreting feedback often involves multidimensional commentary on various aspects of performance, including accuracy and completeness of information, cohesion and coherence, delivery, non-verbal communication, interpreting strategy use and errors (Balaman, 2024; Holewik, 2020; Lee, 2015; Tahraoui, 2022).

While formative feedback is instrumental in enhancing learner's reflection and autonomy (Crezee; Grant, 2016), its provision is often time-consuming and demanding for teachers due to expectations of timeliness, continuity, and depth (Han; Lu; Fan, 2025). As a result, automated feedback systems, particularly those powered by LLMs, have begun to emerge in various pedagogical contexts, such as programming instruction, second-language writing, and translation training (Pankiewicz; Baker, 2023; Teng, 2024; Xu; Su; Liu, 2025). However, studies on LLM-generated feedback in interpreting pedagogy remain scarce. This new feedback mechanism has been praised for its strengths in detail, fluency, and coherence, making it a potentially valuable pedagogical resource (Teng, 2024). However, several studies have also highlighted important challenges. LLMs' lack of contextual awareness, limited sensitivity to learners' individual needs may reduce the relevance and pedagogical effectiveness of their feedback (Guo; Wang, 2024). Moreover, issues such as redundancy and repetition in content can negatively impact student engagement (Chen *et al*., 2024). In a comparative study on engagement strategies, Yu, Wei and Chen (2025) also found that interpreting trainees were more likely to dispute with the AI-generated feedback than peer feedback, indicating lower perceived credibility.

These findings suggest that although AI-powered feedback systems hold considerable potential, their pedagogical value depends heavily on the quality and contextual appropriateness of the feedback they provide. Furthermore, broader educational and ethical consequences of such formative feedback must be addressed, including the risk of bias and the equitable accessibility of these technologies for learners from diverse backgrounds (Sun, 2025). A critical and reflective approach to the use of automated feedback systems is therefore essential to ensure their responsible integration in teaching and benefits for student learning.

1.2 Automatic assessment of interpreting quality

Quality assessment underpins both interpreter training and interpreting studies. Existing assessment methods generally fall into two categories: human and automatic, depending on the agent involved. With the advancement of technology, there has been increasing efforts to develop automated approaches to streamline and enhance assessment processes, as noted by Han and Lu (2021) in their critical review.

According to the literature, interpreting quality is typically evaluated along three main dimensions: content accuracy, fluency, and language quality (Han, 2021; Liu 2021). Among these, fluency assessment is particularly challenging due to the oral nature of interpreting, which often requires the identification and extraction of paralinguistic features, a process that has traditionally relied on manual annotation using acoustic analysis tools such as PRAAT or Cool Edit Pro (Wu, 2021; Yu; van Heuven, 2021). To reduce labor demands, Wang and Wang (2024) introduced automated fluency parameters derived from speech recognition data, offering a methodological

solution for the development of machine learning–based systems for automatic interpreting assessment.

In addition to fluency, other quality dimensions have also been explored. Ouyang *et al*. (2021) focused on linguistic and discourse features, utilizing Coh-Metrix, a tool originally developed for text analysis, to automatically extract linguistic indices related to language use. Based on textual analysis, Liu (2021) conducted a corpus-based study, employing a keyword extraction tool to analyze information accuracy by comparing source and target texts, as well as a corpus tool to extract linguistic features for analysis of linguistic performance. Similarly, further attempt involves the introduction of four machine translation evaluation metrics (BLEU, NIST, METEOR, TER) to assess the information accuracy of human interpretation (Han; Lu, 2021). These studies demonstrate the diversity of methodological pathways through which researchers have sought to conduct automatic assessment of interpreting quality.

With the emergence of LLMs, researchers have begun exploring their potential in both translation and interpreting assessment. In the domain of translation, recent studies demonstrate that LLMs can achieve strong performance in evaluating machine translation outputs, outperforming traditional automatic metrics such as COMET, BLEURT, chrF, and BLEU (Kocmi; Federmann, 2023). Employing error analysis based on the Multidimensional Quality Metrics framework (Freitag *et al*., 2021), LLM-based evaluations have been shown to provide greater interpretability in quality assessment (Lu *et al*., 2023).

While still limited, the application of LLMs to interpreting assessment is gaining traction. Ünlü (2023) tested the feasibility of a self-tutoring interface integrating Whisper (an automatic speech recognition tool) and ChatGPT-3.5/4 for the purpose of English-to-Turkish consecutive interpreting assessment and provision of feedback, though based on only two interpreting samples. Due to the current inability of LLMs to directly process audio input, most study have relied on transcript-based assessment, focusing on dimensions such as information accuracy and linguistic quality. One such effort is the study which investigated the use of GPT and Claude for the automatic assessment of target language quality in English-to-Chinese consecutive interpreting (Wang; Wang, 2025). By comparing human and LLM-generated ratings, the researchers reported a general alignment across the three assessment systems. Similarly, Jia and Aryadoust (2024) prompted ChatGPT-4 with a scoring rubric to assess the Chinese-to-English consecutive interpreting outputs exclusively on information completeness, reporting moderate alignment between the model and human raters in average scores. Expanding beyond single-dimension assessment, Han, Lu and Fan (2025) preserved disfluency markers in Chinese-to-English consecutive interpreting transcripts to enable GPT-3.5 to assess fluency, alongside information completeness and target language quality. Their study demonstrated strong correlations with human ratings and suggested broader potential for LLM-based interpreting assessment. Despite these promising developments, the exploration of LLMs for interpreting quality assessment remains limited. Existing studies tend to target a single quality

dimension and focus predominantly on Chinese–English consecutive interpreting, leaving simultaneous mode and other language pairs largely unexplored.

2. THE PRESENT STUDY

Building on recent advances in LLM-assisted approaches to interpreting quality assessment, the present study explores the use of ChatGPT, a state-of-the-art LLM-based chatbot, in the context of Chinese-to-Portuguese simultaneous interpreting training. Specifically, it examines the feasibility of using ChatGPT to support formative assessment by generating criterion-referenced quality scores and diagnostic feedback based on the full interpretation transcripts. Given that LLMs currently lack the capacity to directly process audio input and therefore cannot evaluate paralinguistic features such as intonation and silent pauses, phenomena that are salient in simultaneous interpreting, this study focuses on two dimensions: information completeness and target language quality of interpreting outputs.

The research is guided by the following questions:

RQ (1): Can ChatGPT generate valid quality scores for Chinese-to-Portuguese simultaneous interpreting?

RQ (2): How do human raters perceive the quality of the diagnostic feedback generated by ChatGPT?

3. METHODS

To address the research questions, we adopted a primarily quantitative design, supplemented by limited qualitative input. The study draws on multiple data types, including interpreting outputs from interpreters with varying levels of proficiency, along with corresponding human ratings to serve as a golden standard (Kocmi; Federmann, 2023; Lu *et al*., 2023). At the same time, ChatGPT was applied to the same outputs, functioning as a formative assessment system to generate quality scores and diagnostic feedback. This section describes the research tools, analytic approaches, and ethical considerations.

3.1 Research tools and implementation

This section introduces the four research tools that together constitute the methodological framework of the study. In this context, “tools” refers to the core methodological resources used to support the study’s data generation and analysis. Table 1 provides an overview of each tool, the stakeholders involved, its primary use and related research questions.

Table 1 – Overview of research tools

Research tool	Participants	Purpose	Related RQ(s)
Interpreting corpus	Professional interpreters Student interpreters	Provides a source-target paired dataset for human and ChatGPT assessment and feedback generation	RQ1 & RQ2
Human rating rubric	Human raters	Defines the assessment criteria and yields human benchmark ratings for comparison with ChatGPT-generated scores	RQ1
ChatGPT prompting protocol	ChatGPT	Elicits quality scores and diagnostic feedback from ChatGPT	RQ1 & RQ2
Post-hoc questionnaire	Human raters	Captures perceptions of the quality of ChatGPT-generated feedback	RQ2

Source: authors

The subsequent subsections detail how the tools were implemented, including stakeholder roles, key procedures, and design rationale.

3.1.1 Interpreting corpus

The interpreting corpus consists of source speeches and the corresponding Chinese-to-Portuguese simultaneous interpreting outputs, together with paired transcripts prepared for analysis. Source speeches and interpretations were transcribed using OpenAI Whisper, an automatic speech recognition system, and then manually corrected for accuracy. Paralinguistic features such as pronunciation errors, repetitions, and self-corrections were preserved in the transcripts to approximate the oral delivery, thus representing the real quality of the interpreting tasks. The paired source and target transcripts together served as the common input for both human and ChatGPT-based rating and feedback generation.

To be specific, the corpus includes 60 authentic Chinese-to-Portuguese simultaneous interpreting tasks produced by 22 student interpreters from two universities in Macau, as well as by three professional interpreters with over five years of work experience. Based on their training experience, the student interpreters were further categorized into two groups: beginners (n=10, with no prior training or no more than six months of training, and no work experience) and advanced trainees (n=12, with over six months of training). Participants also differed in their linguistic backgrounds, with the majority being native speakers of Mandarin and a small subset speaking Cantonese, reflecting the multilingual context of Macau. The inclusion of participants with diverse backgrounds allows us to examine whether the assessment system can fairly assess interpreting quality across different levels of proficiency and language backgrounds.

The 60 interpreting tasks are based on 20 different original speeches in Chinese, each ranging in length from approximately 3 to 6.5 minutes and varying in difficulty. All speeches were either drawn from publicly available online source or prepared by trainers and trainees for use in classroom-based interpreter training. The topics are diverse, covering both general topics, such as opinions on euthanasia, artificial intelligence, and expression skills, and specialized domains, including medicine, stock, and bond markets. This thematic variety allows for an examination of ChatGPT's capacity to assess interpreting quality across a wide range of content types.

In summary, the diversity of both interpreter profiles and source materials in the interpreting corpus provides a solid basis for evaluating ChatGPT as a rating tool and feedback provider, supporting the robustness of the findings within the scope of the present dataset.

3.1.2 Human rating rubric

Human ratings were used to capture the quality of the collected interpreting outputs and served as the benchmark against which the validity of ChatGPT-generated scores was examined.

For this performance-based assessment task, the rubric-referenced rating scale developed by Han (2021) was applied, which comprises three dimensions of interpreting quality: information completeness, target language quality, and fluency. Human raters assessed the interpreting output using both the source and target speech transcripts, as well as the audio recordings of the interpretations, consistent with common authentic assessment practice. In contrast, ChatGPT was provided only with transcripts data, as the model cannot directly process audio. Moreover, transcripts cannot fully capture prosodic features of the interpreting outputs. For these two reasons, only the first two dimensions were assessed in both pathways.

Previous studies have involved various types of human raters, including professional interpreters, interpreting students, translation teachers, and even students without prior interpreting training (Han, 2018). In this study, nine raters (one male, eight female) participated in the human assessment. They were all doctoral students in Portuguese with prior interpreting training experience and were also prospective interpreting teachers. This background enabled them to provide two distinct perspectives: those of interpreting trainees and of future trainers.

Given the relatively large number of interpreting outputs to be evaluated, the nine raters were randomly assigned into three groups (A, B, and C) to reduce the potential impact of rater fatigue on scoring accuracy. Each group consisted of three raters, and each rater was responsible for assessing the same set of 20 interpreting tasks. In total, the assessment covered 60 tasks distributed evenly across the three groups (20*3). Prior to the assessment, the raters participated in a training session aimed at familiarizing them with the assessment criteria and reducing potential inconsistencies in scoring. The training process included a presentation of the rubric, a practice session in which the raters assessed three sample exercises, and a subsequent discussion session

where they compared scores and exchanged views to achieve calibration. Once consensus was reached at the end of the training, the raters independently assessed their assigned 20 interpreting exercises.

3.1.3 ChatGPT prompting protocol

ChatGPT was prompted to produce quality scores and diagnostic feedback for each interpreting task, which constituted the model-based assessment in this study. To ensure comparability with the human pathway, ChatGPT assessed the same set of interpreting outputs and was instructed to follow the same rating rubric. Due to the model's limitation in audio input processing, all prompts were based solely on the textual transcripts of the original and interpreted speeches.

This study employed the web-based version of ChatGPT-4o, the latest model in the GPT family at the time when the study was conducted. This choice was motivated by the aim of the present research, which was to explore the potential of the model for interpreting assessment and feedback generation in educational contexts. In real-world interpreter training, students typically access ChatGPT through its web interface rather than via an API. Using the web-based version therefore allowed us to test a readily accessible and user-friendly configuration, providing a practical paradigm for integrating AI-assisted learning into interpreter education.

Effective use of LLMs for interpreting quality assessment and feedback generation requires carefully designed prompts, which generally include elements such as task context, explicit instructions, and an expected output format (Korzynski *et al*., 2023; Zheng *et al*., 2023). In the present study, the prompt consisted of these three main components: (1) a role definition, assigning the model the role of an experienced interpreting trainer; (2) explicit assessment instructions, explaining that the model should use the provided rubric to assign scores for each interpreting product in terms of information completeness and target language quality; and (3) an output format, specifying numerical scores, as well as explanatory comments that highlight both strengths and errors, thereby constituting diagnostic feedback **(see Appendix)**. Both the prompt and the feedback were written in Chinese, as this is the native language of the intended end-users of the feedback. All automated scoring and feedback generation were carried out on 3 July 2025.

3.1.4 Post-hoc questionnaire

A *post-hoc* questionnaire was administered to the nine raters involved in the rating process to capture their subjective perceptions of the ChatGPT-generated feedback. As noted earlier, these raters were able to provide valuable insights from two distinct perspectives: those of interpreting trainees and of potential future interpreting trainers.

The questionnaire was developed to examine user perceptions across five dimensions. First, drawing on the design of Nazaretsky *et al*. (2024), content objectivity (represented by the terms

"precise", "fair", and "factual") and usefulness (represented by the terms "relevant", "informative", and "applicable") were assessed, with raters indicating agreement with the statement "*I think this feedback is [term]*". Second, given that one of the ChatGPT's roles in this study was an automated feedback tool, its usability was also examined. This focus is consistent with Davis, Bagozzi and Warshaw's (1989) Technology Acceptance Model, which posits that perceived usability and usefulness jointly shape users' intention to use a technology and, ultimately, its actual use. Accordingly, two items were included to address these practical considerations: perceived ease of use and the likelihood of future adoption in interpreting education. Third, given the increasing interest in AI-human feedback integration in various educational contexts (Er *et al*., 2024; Guo; Wang, 2024), the questionnaire also probed user preferences, whether for teacher, AI, or hybrid feedback, as well as the perceived role of AI-generated assessment feedback in interpreter training. Finally, recognizing that assessment feedback can have an emotional impact that may influence learners' willingness to engage with it (Xu; Su; Liu, 2025), the questionnaire also included items to explore both positive and negative emotions elicited by the feedback, as well as its potential effect on the student's motivation.

In total, the questionnaire contained 16 Likert-scale items measuring the five dimensions (1 = strongly disagree, 5 = strongly agree), and one open-ended question inviting additional comments and suggestions on the feedback. It was administered online after the raters completed their assessment tasks, in Microsoft Word format, along with a complete ChatGPT-generated feedback sample corresponding to one interpreting task they had rated.

3.2 Data analysis procedures

Based on the implementation of the research tools described above, multiple types of data were available for analysis, including human ratings, ChatGPT-generated scores and feedback, and human raters' perception data.

We first examined the consistency among human raters to establish a reliable benchmark for subsequent analysis. To this end, both the intraclass correlation coefficient [ICC (3,k)] and Kendall's coefficient of concordance (*W*) were calculated for each assessment dimension, based on the ratings provided by the three raters within each group. The resulting inter-rater reliability coefficients are reported in Section 4.1

Following this step, we compared human ratings with the corresponding ChatGPT-generated scores for the same interpreting tasks. For each task, scores were obtained from both sources on two dimensions, namely, information completeness and target language quality. The comparative analyses allow to examine the validity of ChatGPT-based assessment, with all analyses conducted separately for the two dimensions.

According to Messick's (1995) unified validity framework, validity encompasses six aspects: content, substantive, structural, generalizability, external, and consequential. In the present study, evidence for the first three aspects was ensured by the adoption of Han's (2021) empirically validated rubric (Jia; Aryadoust, 2024; Han; Lu, 2021), applied consistently across both human and automated assessment.

To address generalizability, the dataset included interpreting outputs from participants with varying proficiency levels and across a range of topics (see Section 3.1.1), thereby enhancing the generalizability of the findings. In addition, know-group validity was examined by grouping performances into four quality bands (as predefined in the rubric) based on human ratings, and testing whether ChatGPT scores could significantly distinguish among them using one-way ANOVA. Furthermore, regression analyses were performed to assess the predictive validity of ChatGPT scores with respect to human-rated interpreting quality. For external aspect, which involves convergent and discriminant evidence from multimethod comparisons (Messick, 1995), the study examined concurrent validity by calculating Pearson correlation coefficients between human and ChatGPT scores. Potential systematic bias was also examined via one-sample t-test on the mean score differences. Regarding the consequential aspect, relevant questionnaire items were analyzed to capture raters' perceptions of the potential pedagogical and affective consequences of adopting such ChatGPT-generated feedback in interpreter training.

The rating-based quantitative analyses described above, with results reported in Section 4.2, provide an objective evaluation of the validity of ChatGPT as an interpreting assessment system. We then analyzed the *post-hoc* questionnaire data to capture raters' subjective perceptions of the quality of ChatGPT-generated feedback. Internal consistency of the questionnaire was examined using Cronbach's alpha and the item-level descriptive statistics (means and standard deviations) were also calculated (reported in Section 4.3) to summarize raters' perceptions across different aspects of feedback. In addition, we analyzed responses to an open-ended questionnaire item to elicit raters' further views and suggestions. Taken together, these complementary perspectives provide converging evidence for the feasibility of integrating ChatGPT into interpreter education.

3.3 Ethical considerations

Prior to data collection, informed consent was obtained from all participants, including 25 interpreters (22 student interpreters and 3 professional interpreters) and 9 raters. Each participant received a consent form detailing the study's purpose, procedures, data usage, and their rights.

Once received by the researchers, all simultaneous interpreting recordings were anonymized. Identifiers such as "BG1" (Beginner 1) and "TR1" (Trainee 1) were used to replace names. Additionally, care was taken to ensure that raters did not personally know the interpreters, reducing the possibility of speaker identification based on voice.

Participation in the study was entirely voluntary. Student interpreters' performance was not linked to any course assessment. Raters' evaluations were used solely for research purposes. All audio and rating data were securely stored on password-protected devices, accessible only to the research team.

4. RESULTS

This section reports the results of the analysis described in Section 3.2 to address the two research questions.

4.1 Reliability of human ratings

Given that human ratings served as the benchmark for the comparative analyses, it was essential to establish their inter-rater reliability prior to the validity analyses. Three rater groups (A, B, and C) independently assessed the interpretations' quality, and reliability was calculated separately for the two quality dimensions.

Table 2 – Inter-rater reliability results

	Information completeness (InforCom)				Target language quality (TLQual)			
	Kendall's *W*:	*p*-value	ICC (3,k)	*p*-value	Kendall's *W*:	*p*-value	ICC (3,k)	*p*-value
Group A	0.69	.004**	0.78	< .001***	0.65	.007**	0.79	< .001***
Group B	0.84	< .001***	0.91	< .001***	0.70	.003**	0.80	< .001***
Group C	0.85	< .001***	0.92	< .001***	0.78	< .001***	0.87	< .001***

Source: authors

Note: Interpretation thresholds: for *W*: 0.60–0.79 = good agreement; ≥0.80 = excellent agreement (Gisev et al., 2013); for ICC: 0.75–0.90 = good reliability; >0.90 = excellent reliability (Koo & Li, 2016). Significance levels: $p < .05$ (*), $p < .01$ (**), $p < .001$ (***).

As shown in Table 2, both the *W* values and ICC (3,k) coefficients indicated good to excellent agreement across all groups, on both dimensions. The ICC (3, k) results further confirmed the reliability of mean scores within each group, providing a strong justification for their use in subsequent comparisons with ChatGPT scores.

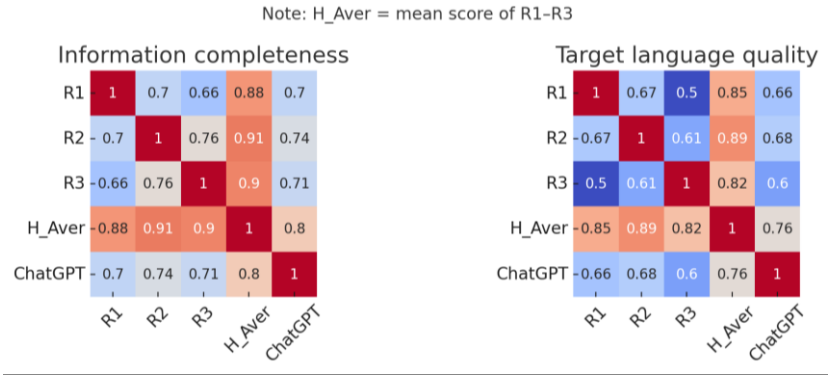
4.2 Validity of ChatGPT-based automatic assessment

To evaluate the validity of ChatGPT-based assessment, as described in Section 3.2, four types of evidence were examined: concurrent validity, predictive validity, known-group validity, and the presence of any systematic bias.

Prior to these analyses, the distributions of both human and ChatGPT scores were examined to determine the appropriate statistical methods. Normality was assessed using the Shapiro–Wilk test and descriptive statistics. For the human benchmark scores (mean scores), both information completeness and target language quality dimensions met the assumption of normality ($p > .05$). For the ChatGPT scores, the Shapiro–Wilk results indicated a statistically significant departure from normality ($p < .001$) for both dimensions; however, skewness and kurtosis values—information completeness (–0.12, –1.26) and target language quality (0.21, –1.17)—fell within the acceptable range of ±2.0 (George; Mallery, 2019), suggesting no substantial deviation from normality. Therefore, parametric statistical approaches were employed in the subsequent analyses.

Concurrent validity was examined by calculating Pearson correlation coefficients between the ChatGPT-based scores and the human benchmark scores. As shown in Figure 1, the correlations between the model scores and human average scores were strong for both information completeness dimension ($r = 0.80$) and target language quality dimension ($r = 0.76$), indicating a high degree of agreement between ChatGPT and human ratings. The correlation was slightly higher for information completeness, suggesting that ChatGPT's scoring in this dimension more closely aligned with human assessment.

Figure 1 – Pearson correlation results for concurrent validity of ChatGPT scoring



Source: authors

Table 3 – Regression analysis results for predictive validity of ChatGPT scoring

Dimension	Intercept (SE)	Slope (SE)	*t* value	*p* value	R^2	Residual SE
InforCom	1.55 (0.30)	0.69 (0.07)	10.21	< .001***	0.643	0.902
TLQual	2.10 (0.27)	0.66 (0.07)	8.80	< .001***	0.572	0.944

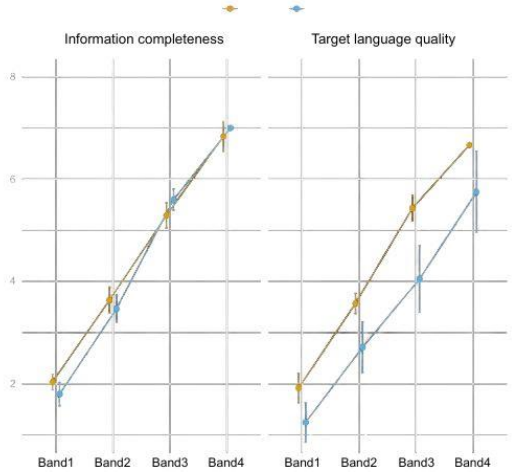
Source: authors

To examine the predictive validity of ChatGPT-based assessment, simple linear regression models were fitted with human mean scores as the dependent variable and ChatGPT scores as the predictor. As Table 3 indicated, in both dimensions, ChatGPT scores significantly predicted human mean scores ($p < .001$). For information completeness, the model explained 64.3% of the variance in human ratings ($R^2 = .643$), whereas for target language quality, the explained variance was 57.2%

(R^2 = .572). According to Ozili (2023), R^2 values above 0.50 are generally considered acceptable in social science research, particularly when predictors are statistically significant. The regression slopes further indicate a stable linear relationship between ChatGPT and human ratings in these two dimensions, supporting the predictive validity of the automatic assessment system.

In terms of the known-group validity of the ChatGPT-based assessment, we followed the band classification defined in the rubric employed during assessment, which evenly categorizes interpreting quality scores (ranging from 1 to 8) into four bands. Given that the participants in this study differed not only in proficiency levels but also in the difficulty of the interpreting tasks they performed, we did not use the original profile-based grouping (i.e., beginners, advanced trainees, and professional interpreters) for this analysis. Instead, to control for these variations, all 60 interpreting outputs were grouped into four bands based on their human mean scores in each assessment dimension. The resulting distribution of human and ChatGPT scores across the four bands is illustrates in Figure 2 (points represent group means, and vertical bars indicate 95% confidence intervals). The visual patterns demonstrate that ChatGPT scores increase in a clear upward trajectory from the lowest to the highest band in both dimensions, closely tracking the human-defined quality levels, which suggests high discriminative capability. In the information completeness dimension, ChatGPT's band-wise scores closely mirror the human ratings. In contrast, for target language quality, ChatGPT's scores are generally slightly lower than the human scores and exhibit greater variability, particularly at the highest quality band.

Figure 2 – Band-wise scores (±95% CI), Human vs ChatGPT



Source: authors

Table 4 – One-way ANOVA results for known-group validity of ChatGPT scoring

Dimension	df	*F*	*p* value	η^2 (partial)
InforCom	3, 56	30.91	6.39e-12 ***	0.623
TLQual	3, 56	17.43	4.1e-08***	0.483

Source: authors

To statistically confirm these observations, a one-way ANOVA was conducted to test whether ChatGPT scores significantly differed across human-defined bands. As shown in Table 4, the results revealed statistically significant differences in both dimensions (all *p*-values < .001). This empirical evidence demonstrates strong known-group validity for the ChatGPT-based assessment, indicating that the model's automated scores reliably discriminate across quality bands of interpreting output.

Systematic bias was assessed by computing, for each task, the score difference (Δ= ChatGPT−human mean) and testing whether the mean Δ differed from zero with a one-sample t-test. For information completeness, ChatGPT's scores were on average 0.26 points lower than human ratings (mean bias=−0.26), a trend that did not reach the conventional .05 threshold (p = .06). For target language quality, ChatGPT's scores were on average 1.01 points lower (mean bias=−1.01), and this underestimation was highly significant (p=1.468e-09, p < .001). These patterns align with Figure 2: within each human-defined band, the two rating sources show near-alignment in the dimension of information completeness, whereas the model's ratings lie systematically below human ratings in target language quality.

4.3 Human perceptions of ChatGPT-generated feedback

Following the scoring phase, raters evaluated the ChatGPT-generated feedback on a five-point Likert scale with respect to five dimensions: content objectivity, content usefulness, usability, user preference, and emotional impact. The questionnaire showed good internal consistency (Cronbach's α = .803).

As summarized in Table 5, content objectivity was rated positively overall (category mean= 3.89), and content usefulness was likewise favorable (category mean=3.78). Perceived usability was strong, especially intention to use (Q8, M=4.78) in future teaching/learning. For preference, raters did not view the AI feedback as a full substitute, yet they strongly favored a complementary role for AI feedback (Q11, M=4.67) and favored combining human and automated feedback (Q12, M=4.89). For emotional impact, negative effects were low, whereas positive impacts were high to moderate (Q15 helps improvement, M=4.56; Q16 increases motivation, M=3.89).

Figure 3: *Post-hoc* questionnaire results for perceived ChatGPT-generated feedback quality

Categories	Items	M	SD
Content objectivity	Q1: I think this feedback is precise.	3.44	0.53
	Q2: I think this feedback is fair.	4.56	0.28
	Q3: I think this feedback is factual.	3.67	1.25
		Average: 3.89	
Content usefulness	Q4: I think this feedback is relevant.	4.22	1.20
	Q5: I think this feedback is informative.	3.44	0.78
	Q6: I think this feedback is applicable.	3.67	1.00
		Average: 3.78	
Usability	Q7: I consider the automated feedback system easy to access and use.	3.89	1.11
	Q8: I would use this automated feedback system in future interpreting teaching and learning.	4.78	0.19
User preference	Q9: I think the feedback of this automated system (ChatGPT) is comparable in terms of quality to human feedback I usually receive	3.00	0.50
	Q10: I prefer to receive human feedback from interpreting trainers than this ChatGPT-generated feedback.	4.11	1.86
	Q11: I think this feedback can serve as a useful complement to human feedback.	4.67	0.25
	Q12: I prefer to accept both human and this automatic feedback, rather than relying on just one or the other.	4.89	0.11
Emotional impact	Q13: I think this feedback triggers anxiety or other negative emotions in students.	1.89	1.11
	Q14: I think this feedback diminishes students' motivation for practices in the future.	1.67	1.00
	Q15: I think this feedback helps students to improve.	4.56	0.28
	Q16: I think this feedback increases students' motivation for practices in the future.	3.89	1.11

Notes: items were rated on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree).

Source: authors

Overall, raters perceived the feedback as fair, usable, and practically valuable, preferred it as a complement to human feedback, and did not report adverse emotional effects. More detailed

analyses and triangulation with responses to the open-ended question are presented in the discussion Section 5.2.

5. DISCUSSION

Building on the Results, this section interprets the patterns revealed by the data, and relates them to our two research questions.

5.1 Potentials and pitfalls of LLM-based interpreting assessment

The findings presented in Section 4.2 provide empirical support for the validity of ChatGPT-based quality assessment of Chinese-to-Portuguese simultaneous interpreting. Specifically, the model's ratings were highly correlated with those of human raters, predictive of human scoring, and able to distinguish between interpreting outputs of differing proficiency levels, thereby supporting its use as a valid assessment system in this context.

A closer look at the two quality dimensions reveals important nuances. In terms of information completeness, ChatGPT outperformed earlier versions reported in Jia and Aryadoust (2024), potentially due to the use of GPT-4o in this study as opposed to GPT-4. According to OpenAI[1], GPT-4o demonstrates significantly improved performance on non-English languages. Given that this study involves both Chinese and Portuguese comprehension and generation, it is plausible that the model's improvements contributed to the higher alignment with human judgments in this dimension.

However, for target language quality, the model consistently underrated participants' performance, with an average difference of 1.01 points compared to human scores. This conservative bias aligns with prior findings in studies focusing on English-Chinese consecutive interpreting (Han; Lu; Fan, 2025; Wang; Wang, 2025). One possible explanation is that LLMs are trained predominantly on written text corpora (Brown *et al*., 2020) and may therefore penalize spoken-language features, such as repetitions, false starts, or self-repairs, that are typical of spoken discourse but deviate from written norms. Such features are more common in simultaneous interpreting, where time pressure increases the likelihood of incomplete sentences and false starts (Gile, 2001). In contrast, the consecutive mode allows more time for contextual analysis and message planning, and interpreters may benefit from note-taking to improve fluency (Zhou; Dong, 2024). Thus, the prevalence of disfluency in simultaneous interpreting may lead to a higher risk of penalization. Although this hypothesis was not directly tested in the present study, it raises important questions for future research, regarding the interaction between interpreting mode, disfluency patterns, language pairs, and the accuracy of LLM-based assessment.

[1] https://openai.com/index/hello-gpt-4o/

This observed discrepancy in the assessment of target language quality may also be rooted in differences in the underlying evaluation logic. Human raters typically adopt a more integrative and pragmatic view of interpreting quality, recognizing the communicative and cognitive demands of the task (Wang; Wang, 2025). In contrast, despite receiving the same multidimensional assessment rubric as human raters, the opaque nature of LLM decision-making precludes any definitive understanding of whether ChatGPT balances these dimensions in a manner comparable to human raters. Its generated feedbacks also suggest a tendency to disproportionately penalize deviations from grammatical correctness or formal language use, even when such issues do not impede information transfer. This scoring tendency may affect how learners perceive their own performance, potentially leading them to overemphasize surface-level language features. It is therefore advisable that future implementations of LLM-based assessment incorporate mechanisms to distinguish between critical and non-critical linguistic issues in order to better align automated assessment feedback with pedagogical priorities.

From the prompting perspective, prior studies typically adopted sentence-level prompting for LLM assessment, while human raters assessed performance at the text level (Han; Lu; Fan, 2025; Wang; Wang, 2025). By contrast, the present study provides both human raters and ChatGPT with full transcripts of the simultaneous interpreting tasks, which may have allowed the model to better capture global coherence and adequacy of information transfer. This approach aligns more closely with the nature of simultaneous interpreting, where sentence-by-sentence equivalence is rarely achievable due to the adoption of interpreting strategies such as omission of redundancy and minor details, simplification, and restructuring (Gieshoff; Albl-Mikasa, 2024). Importantly, Wang and Fantinuoli (2024) found that increasing context window size improves the correlation between LLM and human ratings, further supporting the value of text-level analysis. As for sentence-level assessment, it offers the advantage of fine-grained evaluation and potentially greater scoring transparency and interpretability, as total scores can be averaged from localized judgments. However, it also presents limitations: it requires labor-intensive segmentation and precise alignment with source text, reducing its scalability in real-world educational settings.

These observations highlight the potential of LLMs as effective tools for interpreting assessment, while also revealing areas for further refinement. In particular, the choice of scoring granularity (sentence-level vs. text-level) appears likely to shape the sensitivity of the assessment and thus warrants further validation. In practice, the scalability and pedagogical robustness of LLM-based assessment systems may benefit from technical improvements in transcription accuracy, automatic detection of paralinguistic features, and segmentation tools.

5.2 Insights into the quality and integration of LLM-generated feedback

In addition to evaluating the assessment scores, this study also examined human raters' perceptions of the ChatGPT-generated feedback. Questionnaire results indicated generally positive responses regarding the objectivity and usefulness of the feedback content. However, scores on the "precise" and "factual" sub-dimensions of objectivity, triangulated with responses to the open-ended questions, revealed some concerns. Several raters noted the absence of commentary on fluency, a crucial component of interpreting quality, thus questioning whether the feedback could be considered fully factual. Others found the feedback insufficiently precise, as it cited only some examples of errors rather than comprehensive coverage. Despite these limitations, most participants considered the feedback highly useful. Although many raters (77.8%) still expressed a preference for human feedback, all acknowledged that AI-generated feedback could serve as a valuable supplementary resource in interpreter education and endorsed a hybrid feedback model, combining human and LLM-generated comments, as the most promising approach for instructional integration.

In terms of emotional impact, participants generally believed that such diagnostic feedback could enhance learner motivation and progress. Given the interpreting direction, from Chinese (L1) to Portuguese (L2), many beginner-level student interpreters struggled with target language expression, with an average human-assigned score of only 2.9 out of 8 for language quality. In this context, raters found the linguistic suggestions in the feedback especially helpful for students' lexical development and for reducing language interference. For instance, when a student attempted to express the concept of "facilities", likely influenced by English, the student produced "as facilitações" ([the facilitations]) in Portuguese. The feedback addressed this issue: "*'as facilitações' 应为 'as infraestruturas'*" (["facilitations" should be "infrastructures"]). In another case, the model not only corrected the misuse of the word, but also provided synonym suggestions: "*'difusão' 不明所指，应为 'confusão', 'desordem'*" (["diffusion" is unclear and should be "confusion" or "disorder"]). These examples reflect the model's potential to assist learners in refining their lexical choices and improving the clarity and precision of their target language production. However, while this feedback was generally appreciated, some raters pointed out that it lacked breadth and detail. As one rater explicitly noted in the open-ended question: "*I prefer corrective feedback. I think this feedback only pointed to some errors in the interpretation, but I hope the error correction could be more detailed and this feedback could include more reformulations or usable sentence patterns*." This suggests that LLMs' linguistic capabilities could be more fully leveraged to offer more comprehensive corrective suggestions and multiple phrasing options, thereby helping learners enhance the richness and flexibility of their linguistic resources for interlingual message transfer.

Only one rater raised concerns about the directness of some evaluative comments, pointing out that comments such as "multiple grammatical errors" or "seriously impairs overall language quality and comprehensibility" might cause anxiety among less confident learners. This highlights the importance of considering affective impacts of feedback when designing prompts. For instance,

LLMs could be prompted to adopt a more learner-sensitive tone by first acknowledging the strengths of the performance before offering constructive criticism.

In general, questionnaire results indicated that raters held positive views regarding the quality and usability of the LLM-generated assessment feedback. Such favorable reception is important for the formative function of feedback: when learners value feedback, they are more likely to engage with it, make evaluative judgement and translate it into improved performance (Carless; Boud, 2018). To promote sustained learning benefits, it is recommended to embed the ChatGPT-generated feedback within regular instruction and learning. The feedback designed in this study can support formative assessment and serve as an accessible learning resource throughout the entire learning process. With regular use, students may gradually build a personal feedback corpus, allowing for longitudinal tracking of individual progress and skill development. Moreover, as the feedback is automatically generated by an LLM-based chatbot, it significantly reduces teacher workload while enhancing accessibility for a broader range of learners.

However, it is essential to consider the extent to which learners from diverse backgrounds, including those with varying linguistic proficiency, cultural norms, and technological access, can benefit equally from such tools (Sun, 2025). This represents a key ethical concern in the implementation of AI-powered feedback systems. The present findings should therefore be interpreted with attention to potential differences across student populations. To enhance the pedagogical value of LLM-generated feedback in interpreter training, it may be beneficial to provide learners with guidance on understanding feedback, crafting effective prompts, and engaging productively with the feedback system.

CONCLUSION

This study provides empirical evidence that ChatGPT can generate valid, rubric-referenced quality scores for Chinese-to-Portuguese simultaneous interpreting and can produce diagnostic feedback that human raters perceive as accurate and pedagogically relevant. These findings suggest that, when properly prompted, ChatGPT can support formative assessment in interpreter training. Despite these promising results, research on LLM-based interpreting assessment remains limited. Technological challenges such as audio-based input processing and automatic extraction of paralinguistic features still require further advancement to support more comprehensive and context-sensitive assessment. Moreover, while feedback was generally perceived as useful by human raters, how it can be effectively integrated into interpreting teaching warrant closer investigation. This study also has several limitations. The interpreting dataset was relatively small, and due to accessibility considerations, we used ChatGPT via its web-based interface, which constrained model control and interpretability in a black-box setting. In addition, although the scoring and diagnostic feedback were designed for formative use, the present study did not directly examine the formative effects of using

this feedback. Longitudinal studies that track learners' feedback uptake and performance trajectories would help address this gap and yield further evidence to refine the model's scoring accuracy and feedback quality. Future studies may also consider developing an integrated platform that combines automatic transcription, assessment, and feedback delivery, thereby enhancing usability and supporting sustained learning in interpreter education. Finally, as LLM-based tools for assessment gain traction, future research may also examine their validity across leaner groups and contexts, as well as their fairness and ethical consequences.

REFERENCES

ALAHMADI, N. *et al*. The impact of the formative assessment in speaking test on Saudi students' performance. **Arab World English Journal**, Kuala Lumpur, v. 10, n. 1, p. 259-270, 2019.

BALAMAN, S. Exploring Undergraduate Students' Viewpoints on Corrective Feedback Implementations in Interpreting. **Korkut Ata Türkiyat Araştırmaları Dergisi**, Osmaniye, v. 15, p. 994-1011, 2024.

BOUD, D.; SOLER, R. Sustainable assessment revisited. **Assessment & Evaluation in Higher Education**, London, v. 41, n. 3, p. 400-413, 2016.

BROWN, T. *et al*. Language Models are Few-Shot Learners. **Advances in neural information processing systems**, New York, v. 33, p. 1877-1901, 2020.

CARLESS, D.; BOUD, D. The development of student feedback literacy: enabling uptake of feedback. **Assessment & Evaluation in Higher Education**, London, v. 43, n. 8, p. 1315-1325, 2018.

CREZEE, I.; GRANT, L. Thrown in the deep end: Challenges of interpreting informal paramedic language. **Translation & Interpreting**: The International Journal of Translation and Interpreting Research, Sydney, v. 8, n. 2, p. 1-12, 2016.

DAMACENA, M.; QUEVEDO-CAMARGO, G. Avaliação e formação de professores de línguas: uma discussão sobre o currículo e as percepções dos formandos. **Olhares & Trilhas**, Uberlândia, v. 23, n. 3, p. 1054-1073, 2021.,

DAVIS, F. D.; BAGOZZI, R. P.; WARSHAW, P. R. User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. **Management Science**, Catonsville, v. 35, n. 8, p. 982-1003, 1989.

ER, E. *et al.* Assessing student perceptions and use of instructor versus AI-generated feedback. **British Journal of Educational Technology**, London, v. 56, n. 3, p. 1074-1091. 2024.

FOWLER, Y. Formative assessment: Using peer and self-assessment in interpreter training. *In*: WADENSJO, C.; DIMITROVA, B. E.; NILSSON, A. (eds.). **The Critical Link 4**: Professionalisation of interpreting in the community. Amsterdam: John Benjamins, 2007. p. 253-262.

FREITAG, M. *et al*. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. **Transactions of the Association for Computational Linguistics**, Cambridge, v. 9, p. 1460-1474, 2021.

GEORGE, D.; MALLERY, P. **IBM spss statistics 26 step by step**: A simple guide and reference. New York: Routledge, 2019.

GIESHOFF, A. C.; ALBL-MIKASA, M. Interpreting accuracy revisited: a refined approach to interpreting performance analysis. **Perspectives**, London, v. 32, n. 2, p. 210-228, 2024.

GILE, D. Consecutive vs. Simultaneous: Which is more accurate? **Interpretation Studies**: The Journal of the Japan Association for Interpretation Studies, Tokyo, v. 1, p. 8-20, 2001.

GISEV, N.; BELL, J. S.; CHEN, T. F. Interrater agreement and interrater reliability: key concepts, approaches, and applications. **Research in Social and Administrative Pharmacy**, New York, v. 9, n. 3, p. 330-338, 2013.

GLAZER, N. Formative plus Summative Assessment in Large Undergraduate Courses: Why Both? **International Journal of Teaching and Learning in Higher Education**, Fort Collins, v. 26, n. 2, p. 276-286, 2014.

GUO, K.; WANG, D. To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. **Education and Information Technologies**, New York, v. 29, n. 7, p. 8435-8463, 2024.

HAN, C. Using rating scales to assess interpretation: Practices, problems and prospects. **Interpreting**, Amsterdam, v. 20, n. 1, p. 59-95, 2018.

HAN, C. Detecting and measuring rater effects in interpreting assessment: A methodological comparison of classical test theory, generalizability theory, and many-facet rasch measurement. *In*: CHEN, J.; HAN, C. (eds.). **Testing and assessment of interpreting**: Recent developments in China. Singapore: Springer, 2021. p. 85-113.

HAN, C. Interpreting Testing and Assessment: A state-Of-The-Art Review. **Language Testing**, London, v. 39, n. 1, p. 30-55, 2022.

HAN, C.; LU, X. Can automated machine translation evaluation metrics be used to assess students' interpretation in the language learning classroom? **Computer Assisted Language Learning**, London, v. 36, n. 5-6, p. 1064-1087, 2021.

HAN, C.; LU, X. Interpreting quality assessment re-imagined: The synergy between human and machine scoring. **Interpreting and Society**: An Interdisciplinary Journal, Beijing, v. 1, n. 1, p. 70-90, 2021.

HAN, C.; LU, X.; FAN, Q. Taming generative AI for interpreter education: using large language models in classroom-based assessment of English-Chinese consecutive interpreting. **The Interpreter and Translator Trainer**, London, v. 19, n. 3-4, p. 444-464, 2025.

HATTIE, J.; TIMPERLEY, H. The power of feedback. **Review of educational research**, Thousand Oaks, v. 77, n. 1, p. 81-112, 2007.

HOLEWIK, K. Peer feedback and reflective practice in public service interpreter training. **Theory and Practice of Second Language Acquisition**, Katowice, v. 6, n. 2, p. 133-159, 2020.

IMRAN, M.; ALMUSHARRAF., N. Analyzing the role of ChatGPT as a writing assistant at higher education level: a systematic review of the literature. **Contemporary Educational Technology**, Podgorica, v. 15, n. 4, e464, 2023.

JIA, Y.; ARYADOUST, V. The Utility of Generative Artificial Intelligence in Rating Interpreters' Accuracy: A Case Study of ChatGPT-4. *In*: CHAPELLE, C. A.; BECKETT, G. H.; RANALLI, J. (eds.). **Exploring artificial intelligence in applied linguistics**. Ames: Iowa State University Digital Press, 2024. p. 59-72.

KELLY, D. **A Handbook for Translator Trainers**. London: Routledge, 2014.

KOCMI, T.; FEDERMANN, C. Large language models are state-of-the-art evaluators of translation quality. **arXiv**, 2023. arXiv:2302.14520. Disponível em: https://arxiv.org/abs/2302.14520. Acesso em: 17 jan. 2026.

KOO, T. K.; LI, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. **Journal of chiropractic medicine**, Lombard, v. 15, n. 2, p. 155-163, 2016.

KORZYNSKI, P. *et al*. Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. **Entrepreneurial Business and Economics Review**, Kraków, v. 11, n. 3, p. 25-37, 2023.

LEE, J. Feedback on feedback: Guiding student interpreter performance. **Translation & Interpreting:** The International Journal of Translation and Interpreting Research, Sydney, v. 10, n. 1, p. 152-170, 2018.

LIU, Y. Exploring a corpus-based approach to assessing interpreting quality. *In*: CHEN, J.; HAN, C. (eds.). **Testing and assessment of interpreting: Recent developments in China**. Singapore: Springer, 2021. p. 159-178.

LU, Q. *et al.* Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. **arXiv**, 2023. arXiv:2303.13809. Disponível em: https://arxiv.org/abs/2303.13809. Acesso em: 17 jan. 2026.

MESSICK, S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. **American psychologist**, Washington, DC, v. 50, n. 9, p. 741-749, 1995.

NAZARETSKY, T. *et al*. AI or human? Evaluating student feedback perceptions in higher education. *In*: FERREIRA MELLO, R.; RUMMEL, N.; JIVET, I.; PISHTARI, G.; RUIPÉREZ VALIENTE, J. A. (eds.). **Technology Enhanced Learning for Inclusive and Equitable Quality Education:** EC-TEL 2024. Cham: Springer, 2024. p. 284-298.

OUYANG, L. *et al*. Coh-Metrix model-based automatic assessment of interpreting quality. *In*: CHEN, J.; HAN, C. (eds.). **Testing and assessment of interpreting: Recent developments in China**. Singapore: Springer, 2021. p. 179-200.

OZILI, P. K. The acceptable R-square in empirical modelling for social science research. *In*: SALIYA, C. A. (ed.). **Social research methodology and publishing results:** A guide to non-native English speakers. Hershey: IGI global, 2023. p. 134-143.

SAWYER, D. B. **Fundamental aspects of interpreter education:** curriculum and assessment. Amsterdam: John Benjamins Publishing, 2004.

SCARAMUCCI, M. V. R. O professor avaliador: sobre a importância da avaliação na formação do professor de língua estrangeira. *In*: ROTTAVA, L.; SANTOS, S. (orgs.). **Ensino-aprendizagem de línguas:** língua estrangeira. Ijuí: Editora da UNIJUI, 2006. p. 49-64.

SUN, L. Transforming business interpretation education with AI: Perspectives from instructors and learners. **Education and Information Technologies**, New York, n. 30, p. 1-35, 2025.

TAHRAOUI, A. Teaching sight and bilateral interpreting online: students' perceptions of teacher feedback. **Texto Livre**, Belo Horizonte, v. 15, e39545, 2022.

TENG, M. F. “ChatGPT is the companion, not enemies”: EFL learners’ perceptions and experiences in using ChatGPT for feedback in writing. **Computers and Education: Artificial Intelligence**, Oxford, v. 7, e100270, 2024.

ÜNLÜ, C. Interpretutor: Using large language models for interpreter assessment. In: INTERNATIONAL CONFERENCE HUMAN-INFORMED TRANSLATION AND INTERPRETING TECHNOLOGY (HiT-IT), 2023, Naples. **Proceedings** of the International Conference HiT-IT 2023. Shoumen: INCOMA Ltd., 2023. p. 78-96.

WANG, X.; FANTINUOLI, C. Exploring the correlation between human and machine evaluation of simultaneous speech translation. **arXiv**, 2024. arXiv:2406.10091. Disponível em: https://arxiv.org/abs/2406.10091. Acesso em: 17 jan. 2026.

WANG, X.; WANG, B. Identifying fluency parameters for a machine-learning-based automated interpreting assessment system. **Perspectives**, London, v. 32, n. 2, p. 278-294, 2024.

WANG, X.; WANG, B. Advancing automatic assessment of target-language quality in interpreter training with large language models: insights from explainable AI. **The Interpreter and Translator Trainer**, London, v. 19, n. 3-4, p. 465-485, 2025.

WILIAM, D.; THOMPSON, M. Integrating assessment with learning: What will it take to make it work? *In*: DWYER C. (ed.). **The future of assessment**: shaping teaching and learning. New York: Routledge, 2017. p.53-82.

WU, Z. The interrelationship among in-class peer-assessment, interpreting anxiety and interpreting performance. **Language Education**, Abingdon, v. 5, n. 4, p. 33-37, 2017.

WU, Z. Chasing the unicorn? The feasibility of automatic assessment of interpreting fluency. *In*: CHEN, J.; HAN, C. (eds.). **Testing and assessment of interpreting**: Recent developments in China. Singapore: Springer, 2021. p. 143-158.

XU, S.; SU, Y.; LIU, K. Investigating student engagement with AI-driven feedback in translation revision: A mixed-methods study. **Education and Information Technologies**, New York, v. 30, p. 16969-16995, 2025.

XUE, R.; LIU, Q. Exploring student interpreters’ engagement with different sources of feedback on note-taking. **Innovations in Education and Teaching International**, Abingdon, v. 62, n. 4, p. 1135-1148, 2024.

YAN, D.; AMINI, M.; KASUMA, S. A. A. Status quo of the formative assessment enactments in spoken language interpreter training: a scoping review of research and practice. **International Journal of Academic Research in Progressive Education and Development**, Bahawalpur, v. 12, n. 4, p. 652-673, 2023.

YU, W.; VAN HEUVEN, V. J. Quantitative correlates as predictors of judged fluency in consecutive interpreting: Implications for automatic assessment and pedagogy. *In*: CHEN, J.; HAN, C. (eds.), **Testing and assessment of interpreting**: Recent developments in China. Singapore: Springer, 2021. p. 117-142.

YU, Y., WEI, W., CHEN, Z. Comparing learners’ engagement strategies with feedback from a Generative AI chatbot and peers in an interpreter training programme: a quasi-experimental study. **The Interpreter and Translator Trainer**, London, v. 19, n. 3-4, p. 338-356, 2025.

ZHENG, C. *et al*. Progressive-Hint Prompting improves reasoning in large language models. **arXiv**, 2023. arXiv:2304.09797. Disponível em: https://arxiv.org/abs/2304.09797. Acesso em: 17 jan. 2026.

ZHOU, J.; DONG, Y. Effects of note-taking on the accuracy and fluency of consecutive interpreters' immediate free recall of source texts: A three-stage developmental study. **Acta Psychologica**, Amesterdam, v. 248, e104359, 2024.

CADERNOS de LINGUAGEM & SOCIEDADE
Papers on Language and Society

APPENDIX

Prompt used for quality score and diagnostic feedback generation (translated from Chinese):
You are a highly experienced interpreter trainer. Based on the following interpreting quality rating scale (score range: 1–8), please assign scores to the student's interpreting output. According to this scale, interpreting quality should be assessed across three dimensions: information completeness, fluency, and target language quality. Because you will read only the transcripts of the source speech and the interpretation, you only need to assess information completeness and target-language quality.

Please strictly follow the scale descriptions for these two dimensions when scoring. Keep the scoring criteria consistent across all ratings and avoid any shift in style or standards. Provide integer scores only.

Please also provide feedback identifying the student's strengths and errors as your rating rationale.

The interpreting rating scale is as follows:

Band/scoring criteria	Information completeness (InfoCom)	Fluency of delivery (FluDel)	Target language quality (TLQual)
Band 4 (Score range: 7–8)	A substantial amount of original messages delivered (i.e., > 90%), with a few number of deviations, inaccuracies, and minor/major omissions	Delivery on the whole fluent, containing a few disfluencies such as (un)filled pauses, long silence, fillers and/or excessive repairs	Target language idiomatic and on the whole correct, with only a few instances of unnatural expressions and grammatical errors
Band 3 (Score range: 5–6)	Majority of original messages delivered (i.e., 60–70%), with a small number of deviations, inaccuracies, and minor/major omissions	Delivery on the whole generally fluent, containing a small number of disfluencies	Target language generally idiomatic and on the whole mostly correct, with small amount of instances of unnatural expressions and grammatical errors
Band 2 (Score range: 3–4)	About half of original messages delivered (i.e., 40–50%), with many instances of deviations, inaccuracies, and minor/major omissions	Delivery rather fluent. Acceptable, but with regular disfluencies	Target language to a certain degree both idiomatic and correct. Acceptable, but contains many instances of unnatural expressions and grammatical errors
Band 1 (Score range: 1–2)	A small portion of original messages delivered (i.e., < 30%), with frequent occurrences of deviations, inaccuracies, and minor/major omissions, to such a degree that listeners may doubt the integrity of renditions	Delivery lacks fluency. It is frequently hampered by disfluencies, to such a degree that they may impede comprehension	Target language stilted, lacking in idiomaticity, and containing frequent grammatical errors, to such a degree that it may impede comprehension

Source: Han (2021)

O/A(S) AUTOR(ES/AS)

Wenjing Liu

Wenjing Liu is currently PhD candidate in Interpreting Studies at Macao Polytechnic University in China. Her research interests include translation and interpreting quality assessment, interpreter training, and intercultural studies.

Adriana Silvina Pagano

Adriana S. Pagano is Full Professor in Applied Linguistics at Universidade Federal de Minas Gerais, Brazil. She is a research fellow of CNPq (National Council for Scientific and Technological Development, Ministry of Science and Technology, Brazil) and FAPEMIG (Research Foundation of the State of Minas Gerais, Brazil). Her research interests include systemic-functional approaches to translation and interpreting and multilingual and multimodal modelling of meaning.