Language Testing in Asia

RESEARCH Open Access

A proposed analytic rubric for consecutive interpreting assessment: implications for similar contexts

Shilan Shafiei[1*]

*Correspondence: S.Shafiei@scu.ac.ir; Shilan.Shafiei@gmail.com

[1] Department of English Language and Literature, Faculty of Letters & Humanities, Shahid Chamran University of Ahvaz, Ahvaz, Iran

Abstract

The present study aimed to develop an analytic assessment rubric for the consecutive interpreting course in the educational setting in the Iranian academic context. To this end, the general procedure of rubric development, including data preparation, selection, and refinement, was applied. The performance criteria were categorized into content, form, and delivery. Two groups of participants, experts, and students were recruited to establish the rubric's validity and reliability. Based on the statistical analysis, the developed analytic rubric was established as a valid tool for use in the Iranian academic context of consecutive interpreting assessment. The proposed rubric may provide novice trainers with a more objective and systematic tool for consecutive interpreting assessments.

Keywords: Rubric, Analytic rubric, Consecutive interpreting, Assessment

Introduction

Quality assessment in interpreting informs educators and trainees about specific qualifications in both academic and industrial contexts. These practices impact the decision-making and objectives of stakeholders, practitioners, certifiers, and candidates across diverse contexts. Han and Lu (2021) argue that assessments can have far-reaching consequences, influencing stakeholders' professional identity, livelihood, and social accessibility.

Although there are different approaches to scoring, rubric scoring uses a reference scale with detailed descriptions of different performance levels. Rubrics facilitate systematic grading. They encompass all construct sub-components, offering descriptive behavior statements for each. Scoring rubrics enable graders to evaluate all test elements comprehensively (Angelelli, 2009). Descriptors enhance score consistency among independent raters (Moskal, 2019).

Additionally, rubrics offer performance feedback to the assessed, such as interpreting students. Knoch (2007) notes that the advantage of analytical scoring lies in its detailed profiling of students' abilities across sub-traits, which is suitable for diagnostic purposes. Huot (1990) suggests that adding items to a discrete-point test enhances reliability, thus providing multiple scores per text. Reiss (2000) asserts that "developing objective evaluation methods for translations benefits language awareness and critics' linguistic and extra-linguistic understanding" (p. xi). It seems the concept applies to interpreting assessments.

In Iran's academic B.A. programs for English Translation, interpreting is covered through three courses: consecutive interpreting, simultaneous interpreting, and an introduction to interpreting settings. These courses total six credits. However, interpreting still lacks recognition as an autonomous academic discipline. The limited interpreting courses compared to translation and misunderstandings of course objectives have led to chaos in the field (Shafiei et al., 2019).

The absence of a validated assessment tool and limited empirical research implies raters' reliance on impressionistic or individualistic approaches. Consequently, trainers may assess identical interpreted texts differently, leading to divergent results/scores (Shafiei, 2021). A proposed solution to these fluctuations is adopting a scoring framework (Bachman, 1990; McNamara, 1996). Thus, this study aims to advance a more objective rating approach in B.A. consecutive interpreting (CI) courses in Iran and comparable contexts.

Research question

In this study, the following question has been raised:

> Is the newly developed analytic consecutive rubric reliable and valid enough to be used in the Iranian interpreting academic context?

Review of literature

The importance of assessment in educational settings

Assessment is a crucial element of the educational process, enabling educators to evaluate students' skill and knowledge levels while providing valuable feedback on their learning progress. Taras (2005) notes that assessment equips educators with the tools to effectively assess and improve learning outcomes, while Wojtczak (2002) highlights its role in identifying students' strengths and weaknesses and acting as a motivational instrument.

Language assessment, vital in foreign language teaching and learning, typically occurs within language programs. Grounded in the program's content, teachers often design and implement these assessments, incorporating observational techniques, portfolios, self-assessments, and informal and formal tests.

In language teaching, assessment is a key to evaluating student performance and proficiency, employing two fundamental approaches: holistic and analytic rating. Holistic rating, assigning a single score to reflect the overall quality of work, is valued for capturing the rater's immediate reaction to a text. In contrast, analytic rating evaluates multiple criteria separately, providing detailed feedback on aspects such as grammar, vocabulary, coherence, and organization, thereby facilitating tailored instruction.

Holistic scoring is critiqued for its limited diagnostic information (Nelson & Van Meter, 2007). On the other hand, analytical scoring disaggregates performance across various dimensions, avoiding the conflation of different performance aspects into a single score, thus simplifying rater training and enhancing reliability (Knoch, 2009).

Through qualitative research, Kola (2022) demonstrated how technology teachers use analytical rubrics to enhance their teaching by clarifying rubric descriptors and terms, thus effectively guiding students. This emphasizes the importance of clear communication in utilizing analytic rubrics for assessment.

Recent studies have focused on developing and validating analytic rubrics for educational settings. Iriani et al. (2023) developed an analytic rubric for assessing students' abilities in creating objective questions, utilizing the Plomp developmental model. Uludag and McDonough (2022) validated a rubric for evaluating integrated writing in English for academic purposes through mixed methods to establish rubric quality. Similarly, Li (2022) investigated the reliability and internal validity of scoring rubrics in EFL writing assessments.

Given the specificity and diagnostic precision of analytic rubrics, they are precious for identifying detailed aspects of language proficiency, leading to targeted feedback and enhanced instructional strategies; this research aimed to develop an analytical rubric for interpreting assessment in the Iranian academic setting of CI teaching.

Interpreting assessment

Despite the prevalence of interpreting performance assessment in interpreter education, research on the quality of interpreting assessment remains scarce. This gap suggests that assessments rely on intuitive understanding rather than a solid theoretical or empirical foundation (Pöchhacker, 2004; Sawyer, 2004). Struyven et al. (2005) note that clear articulation of assessment criteria can significantly enhance learner autonomy and influence student performance. To bridge this gap, experienced interpreter trainers have developed detailed evaluation sheets to grade students' interpretations.

Accordingly, scholars have devised rubrics tailored to specific interpreting modes and types. Carroll's (1966) rubric, initially for machine-translated texts, has been adapted for interpreting studies (Tiselius, 2009; Anderson, 1994). Pöchhacker (2001) introduced four primary criteria—accurate rendition, adequate target language expression, equivalent intended effect, and successful communicative interaction—that span lexico-semantic to socio-pragmatic aspects. Riccardi (2002) identified 17 micro-criteria for interpreting assessment, including phonological and prosody deviations, pauses, eye contact, and posture. However, Riccardi's (2002) criteria, while applicable for formative assessment, do not offer guidance on translating interpreting quality into numerical scores.

Emphasizing interpreting as an interactive activity, Wadensjö (1998) suggests focusing on the communicative process rather than mere text processing. Bartłomiejczyk (2007) differentiates between external evaluation by trainers and self-evaluation by trainees, the latter serving as a developmental tool. Early work by Russo (1995) explored self-evaluation empirically, aiming to enhance students' awareness of their strengths and weaknesses.

Recent efforts include Lee (2015), who developed an analytic rubric for Korean undergraduate CI trainees, and Bontempo and Hutchinson (2011), who designed a rubric to identify professional interpreters' skill gaps in Australia. Also, Lee (2008) contributed a three-scale analytic rubric for CI assessment. Wang et al. (2015) devoted a part of their study to developing a rubric for sign language interpreting, using four macro-level criteria to evaluate interpreting performance comprehensively.

The following sections review the efforts to develop interpreting assessment rubrics in the local context, followed by an evaluation of the rubrics offered. This evaluation justifies the development of a rubric tailored for assessing CI in the Iranian academic context.

Local rubrics proposed for interpreting assessment

A review of local interpreting assessment literature revealed two rubric examples. Ferdowsi (2014) proposed a skill-based rubric for CI assessment, featuring skills such as 'note-taking,' 'observing TL structure,' and 'coping with different accents.' The rubric categorizes performance into three levels: demonstrating skill, skills not refined, and missing skills without a specified weighting scheme. Emam (2013) developed a rubric based on 'diction,' 'grammar,' 'fluency,' and 'comprehensibility,' allocating a total score of forty without specifying the interpreting mode.

Evaluating the proposed rubrics

Ferdowsi's (2014) rubric

This scale invites several constructive observations for potential enhancement:

- The rationale for selecting specific skills within the rubric could benefit from greater clarity and exposition. The developer should provide information on the basis for selecting these skills. Ferdowsi (2014) solely asserts that "all these skills should be taught during the course at universities and then should be examined at the end of the course to evaluate the number of required skills for each trainee" (p. 411).
- Certain aspects included in the rubric, such as 'volume' and 'pace' of speech, might be more accurately characterized as attributes of a successful performance rather than direct interpretive skills.
- Improving consistency in the language used for writing descriptors would ensure uniformity and clarity across all criteria. For example, consider the following three descriptors: 'ability to cope with different accents of working languages,' 'volume,' and 'note-taking.'
- The rubric presently lacks a distinct scoring system or specific weighting scheme. Implementing a well-defined system for score points might aid in its practical application.
- Some scale descriptors currently present in the rubric are somewhat vague. Making these descriptors more precise would aid raters in providing consistent and accurate ratings. For instance, the criterion 'observing the required strategies for interpreting' could be more explicit about what these strategies entail.
- Incorporating elements of reliability and validity more prominently would strengthen the scale's development process and its overall significance.
- The rubric appears to be primarily intuition-based. A shift towards a more empirically grounded approach could enhance its robustness and applicability.

Notably, the current author interviewed the rubric developer about the aforementioned critical items. The developer confirmed that her rubric is based on intuition and has not been validated for reliability and validity.

Emam's (2013) rubric

While Emam's rubric provides valuable insights, it has invited specific observations that merit further discussion:

- Emam postulates that effective oral communication, and by extension interpreting, hinges on four key elements: diction, grammar, fluency, and comprehensibility. Focusing primarily on oral production, this perspective might seem somewhat narrow when considering the broader spectrum of interpreting, which encompasses a range of definitions and perspectives.
- The researcher highlights important traits for evaluation in interpreting. However, the rationale behind each criterion's assigned weightings appears less articulated. Emam (2013) suggests, "It seems that diction is essential in interpreting evaluation so that the greatest contribution will go for this criterion" (p. 76). Nevertheless, a more detailed justification could enrich the understanding of these weightings.
- The rubric developed by the researcher is intended for use in consecutive and simultaneous interpreting. This approach does not account for the distinct differences inherent in these two modes of interpreting.
- There appears to be a need for a more pronounced focus on reliability and validity, crucial aspects of scale development. These elements seem to have received limited attention in the current framework.
- Like Ferdowsi's rubric, Emam's proposal also seems to lean more towards an intuition-based approach rather than being firmly grounded in empirical evidence.

This study recognizes that further development is needed in these areas and aims to fill the gaps identified in previous research. By capitalizing on the strengths of the analytic assessment method, it endeavors to develop an analytic rubric for CI assessment for B.A. in English Translation, thus contributing to the local sphere of CI assessment literature.

Methodology

The present researcher followed the general procedure of rubric development for undergraduate Korean CI students suggested by Lee (2015). However, modifications were necessary for this independent research project, differing from Lee's (2015) rubric development procedure.

The first stage involved identifying CI criteria through a literature review, including existing local rubrics. To this end, a comprehensive review of existing CI scales outside and inside Iran was conducted to gather rating categories. A list of criteria was then compiled, and the criteria were categorized into three main classifications as identified by Zwischenberger (2010) and previously applied by Lee (2015). The criteria identified and selected from the first stage of the scale development were further modified and refined into clear, well-formed sentences due to the lack of language consistency in the literature's criteria. Then, the criteria were transformed into a questionnaire format and distributed to 20 participants to determine the order of importance of descriptors to assess each criterion's total weightings for the target population and to gauge the content validity. Since there was no existing questionnaire suitable for this study, an instrument was specifically designed, validated, and applied to the study by the present researcher. The questionnaire aimed to address three issues on assessment: (1) the importance of each descriptor in sub-scales, (2) the total weightings of each criterion, and (3) the content validity.

For the first issue, many items for the three main CI assessment criteria were derived from a comprehensive literature review and formulated in sentence format. The second issue addressed the respondents' views on each criterion's contribution to the CI's total performance quality. The final issue was to ensure the content validity of the criteria. The questionnaire utilized a five-Likert scale, and before its implementation**,** it was sent to four experts, three in translation studies and one in linguistics; they subsequently commented on and revised it. Piloting was conducted with three researcher colleagues to establish face validity and finalize the questionnaire. Some parts were adjusted to enhance readability and avoid ambiguity. The researcher employed criterion sampling to reduce bias and achieve more rigorous results (Saldanha & O'Brien, 2013). Only interpreting trainers (with limited ad hoc interpreting experience) who were highly interested in interpreting teaching and research were included in the questionnaire survey. A total of 20 participants completed the questionnaire to determine the weightings of the criteria. The respondents comprised 6 PhD candidates in translation studies, 3 PhD holders, and 11 translation studies M.A. graduates with an average experience of almost 4.5 years in teaching interpreting courses and almost 1.5 years of professional interpreting experience. Finally, in alignment with Lee (2015), the criteria were integrated into the layout of a model rating instrument proposed by Christison and Palmer (2005, as cited in Bachman & Palmer, 2010). However, additional modifications were made to the template.

Moreover**,** a sample of six B.A. English translation trainees participated in a pilot study. They were the researcher's trainees in the interpreting course, both male and female, having completed the same amount of translation courses and one interpreting course. The final exam scores from a prior course were used to evaluate interpreting performance. To ensure the response validity, the researcher elaborated on the participants' assessment criteria, and they engaged in a CI test. A 7-min intermediate-level sociopolitical speech delivered by a native speaker of American English at an average rate of 130 words per minute was used as a CI test, and the participants were asked to interpret the text from English into Persian. The video-recorded data were assessed by two raters using the newly developed analytic rubric. The raters, one of whom was the researcher of this study, both held PhDs in translation studies. They had similar experience in teaching interpreting but lacked professional interpreting experience, except for ad-hoc interpreting. They were both female and in their early forties. In a moderation session, the researcher introduced the assessment tool to her colleague and reviewed the test purpose and assessment criteria. Ethical considerations regarding filmed participants were also discussed. The Pearson correlation coefficient was used to assess the inter-rater reliability, and Cronbach's alpha evaluated the whole scale's internal consistency and the three sub-scales.

Results

The thematic sections are presented and discussed below, following the general data preparation, selection, and refinement procedure in developing the rubric.

Reviewing and collecting the existing criteria

The literature reveals diverse approaches to interpreting quality evaluation. Publications establish various criteria via surveys, real-world simulations, and expert impressionistic views. This research integrated these varied approaches, focusing specifically on CI to gather relevant assessment criteria.

Criteria categorization

Criteria were categorized following Zwischenberger's (2010) framework: content, form, and delivery. This categorization was chosen to provide a structured approach to assessing CI performance.

Problems of criteria categorization

Categorizing the collected criteria presented challenges and pitfalls. Few publications provide a detailed account of these criteria, often offering only general guidelines without thorough operationalization of constructs. A significant drawback of the categorization process was the duplication of specific sub-criteria across different categories. 'Logical cohesion,' for instance, fits both content and form. The frequency informed the researcher's decision-making in such cases of inclusion in literature.

Considering the rubric's intended use by undergraduate students and their trainers, the selection focused on criteria relevant to educational settings. The exclusion of professional standards not pertinent to educational settings underscores the rubric's academic focus and applicability. Thus, sub-criteria like 'thorough preparation of conference documents,' 'endurance,' and 'pleasant appearance'—deemed professional standards by AIIC—were excluded from the CI assessment data. Additionally, 'positive feedback of delegates' was omitted, as it falls under quality assessment from the listener's perspective (Pöchhacker, 2001), warranting separate research.

One issue that needs to be considered is that two other sub-criteria, 'strong memory' and 'strong-note-taking skills,' were deleted. The literature emphasizes that rubric descriptors must be observable; manifestations of a strong memory can be identified through other criteria. 'Strong note-taking skills' were excluded, recognizing that some interpreters, perhaps with good memory, may not use this technique (Shafiei et al., 2017). Therefore, including this criterion was deemed unfair. While note-taking is crucial for accurate message delivery, assessing mastery of note-taking techniques is challenging. Additionally, 'Deixis,' 'Modality,' and 'Speech acts'—discourse elements proposed by Clifford (2001) for professional interpreters—were excluded from the rubric. These elements are beyond the scope of students in the preliminary stages of CI.

Consequently, after addressing challenges in categorization, the final selection of criteria was completed; 45 out of 73 sub-criteria were meticulously finalized and transformed into descriptors. These were subjected to expert validation, an expert in linguistics, and three experts in translation studies for alignment with observable student performance.

Writing the descriptors

The crafting of descriptors was guided by principles of clarity and observability, essential for consistent and reliable assessments. Creating explicit, observable descriptors aligns with Davies et al.'s (1999) emphasis on explicit performance descriptions to minimize rater discrepancies. Wording differences in rubric descriptors can lead to varied interpretations by raters, reduced consensus, and lower reliability. Therefore, in writing standard descriptors, the researcher factored in the intended learning outcomes and described the requirements for students to meet each criterion sufficiently. The researcher aimed for clarity and conciseness in criteria descriptions, ensuring understandability and basing descriptors on observable aspects of student performance. In writing the descriptors, the researcher modified some sub-criteria to make them suitable for inclusion in the rubric. Some of the modifications are as follows:

- Larson (1998) contends that translators should aim for idiomatic, natural receptor language texts that convey meaning rather than adhering closely to the source language's form. Thus, the researcher combined 'natural/idiomatic target-language expressions' and 'minimal source language interference' into 'avoiding literal translation.'
- The terms' sense consistency with the original message,' 'accurate rendition of ideas,' 'equivalent intended effect,' and 'faithful rendering' were consolidated into 'accurate rendering of the source text message in the target text.'
- The criteria' completeness of interpretation,' 'general content,' and 'correct interpretation of source-text propositions' were merged into 'complete rendering of the source text message(s).'

Reliability and validity check

Validity

Reliability and validity checks involved a comprehensive review of interpreting activities and their alignment with the rubric's constructs. Content validity was ensured through expert reviews and focused questionnaires. The rigorous validity check process bolsters the rubric's comprehensive nature, enhancing its academic robustness and practical applicability. The researcher undertook several stages to validate the proposed instrument. Initially, the researcher identified activities involved in interpreting. Sawyer (2004) outlines that interpreters must:

- Interpret with faithfulness to the meaning and intent of the original text.
- Use appropriate language and expression.
- Apply word knowledge and knowledge of the subject.
- Demonstrate acceptable platform skills and resilience to stress.

Tiselius (2009) notes that “valid evidence includes construct and content” (p. 96). “Construct validity, which encompasses all validity types, is the adequacy of a test in measuring the underlying skill” (Gipps, 1994, p. 58). Gipps (1994) asserts that assuring construct validity requires focusing on criteria. Furthermore, McMillan (1997) defines criteria as “clear, public descriptions of student performance facets” (p. 29). Consequently, the study’s second step involved defining and describing criteria for CI interpreting performance. The researcher aimed to clarify the relevant constructs. Operationalizing variables simplifies their use, saving time and effort. Such operationalization broadens the study design’s applicability beyond the studied population. Thus, the researcher elaborated on the criteria and their underlying constructs from the literature. The researcher selected the most relevant to this study despite various existing definitions.

Larson (1998) equates content with ‘meaning,’ and Gile (2009) with ‘information transfer.’ However, how much information is enough and what makes it understandable in each interaction situation remains to be tested. Pöchhacker (2015) notes that fidelity and faithfulness in translation and interpreting are often interpreted as accuracy and completeness in contemporary contexts. Pöchhacker (2015) suggests operationalizing ‘accuracy’ by evaluating error severity in an error deduction approach. Accuracy can be assessed based on the number of correctly rendered propositions in the target language (Liu & Chiu, 2008).

Larson (1998) defines the form of a language as the actual words, phrases, clauses, sentences, and paragraphs used in speech or writing. These are known as the language’s surface structure, evident in both print and speech. In interpreting assessment, form pertains to the rendition’s structure and target language quality. Lee (2008) states that target language quality encompasses linguistic correctness, naturalness, and contextual appropriateness of the rendition.

As Lee (2008) describes, “Delivery involves effective public speaking, presentation, and broader communicative skills” (p. 170). Angelelli (2009) characterizes communication as encompassing Interaction, context, form, gist, gesture, tone, and power dynamics. Interpretation, similar to other communication forms, is multifaceted, involving a sender, channel, and recipient. Glasser asserts that successful communication requires mutual understanding of verbal and non-verbal cues.

The second investigated validity evidence type was content validity, derived from logically and judgmentally analyzing items and instrument format. Consequently, descriptors were formatted into a questionnaire with three questions, including one to assess content validity: “To what extent do you think the descriptors are indicative of the underlying traits involved in assessing the criterion under question?” Thus, three field experts evaluated criteria and descriptors for content and construct validity, aligning with previously mentioned abilities.

The results of validity check

The respondents’ consensus over the questionnaire’s underlying constructs was indicative of the validity of the descriptors. See Fig. 1 for details. The high content validity indicated by respondent consensus reinforces the descriptors’ alignment with interpreting assessment standards.

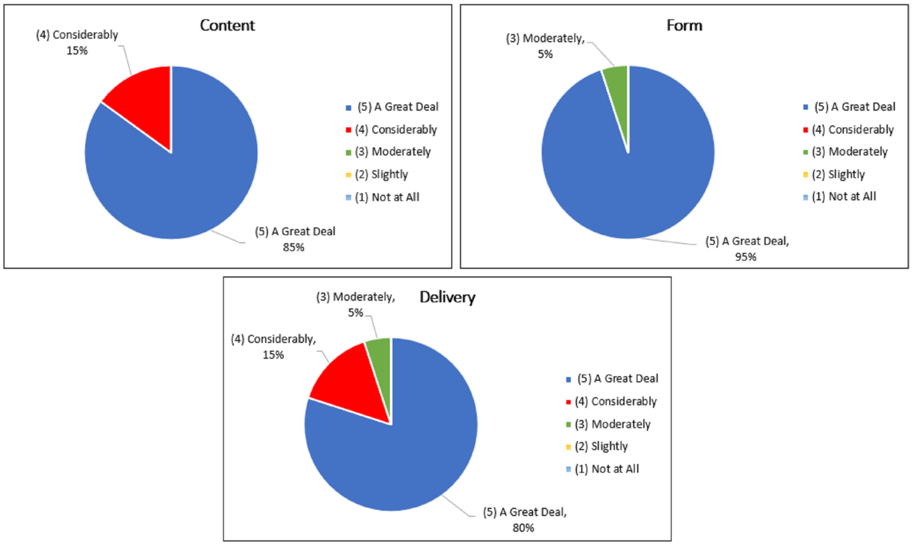

Fig. 1 Results of content validity check in percentage

Table 1 Internal consistency check of the total scale and the three sub-scales

Sub-scale	No. of items	Cronbach's alpha
Total	25	.859
Content	8	.593
Form	8	.731
Delivery	9	.658

The results of the reliability check

The internal consistency of the sub-criteria in the questionnaire was assessed using Cronbach's alpha. The appropriateness, corresponding descriptors, and assigned weights for each criterion were confirmed. See Table 1 for details. The reliability results affirm the rubric's robustness and indicate areas for potential refinement.

Although the standard for an acceptable alpha coefficient is arbitrary and depends on the theoretical knowledge of the scale, alpha coefficients below 0.5 are generally deemed unacceptable. A scale's Cronbach alpha coefficient should ideally exceed 0.7 (Devellis, 2003). However, Pallant (2011) notes that "alpha values are sensitive to the number of items in a scale, and short scales (fewer than ten items) often yield lower values, such as 0.5" (p. 97). "An alpha score above 0.75 indicates high reliability, scores between 0.5 and 0.75 indicate moderate reliability, and scores below 0.5 imply low reliability" (Hinton et al., 2004, p. 363). The present study's scale exhibited high internal consistency, with a Cronbach alpha coefficient of .859. The sub-scale values were .593 for content, .731 for form, and .658 for delivery, indicating moderate reliability.

Table 2 Mean scores obtained for the descriptors of content

Content	1	2	3	4	5	6	7	8
M*	4.4500	4.2000	4.0000	4.4000	3.4000	3.7500	3.3500	3.2500
N**	20	20	20	20	20	20	20	20
SD***	.7592	1.1965	1.2566	.9947	1.4290	1.2513	1.0894	1.0196

*M**Mean, *N***Number of participants, *SD****Std. Derivation

Table 3 Mean scores obtained for the descriptors of form

Form	1	2	3	4	5	6	7	8
M.	3.4000	3.2500	3.6000	3.0500	3.2000	3.1000	3.5000	3.7000
N.	20	20	20	20	20	20	20	20
SD.	1.1877	1.1180	1.3139	1.2763	1.3219	1.2524	1.3179	1.4180

Table 4 Mean scores obtained for the descriptors of delivery

Delivery	1	2	3	4	5	6	7	8	9
M.	3.4500	4.7000	4.7000	3.3500	2.6500	3.4000	3.9500	4.1500	3.3000
N.	20	20	20	20	20	20	20	20	20
SD.	1.0990	.4702	.4702	1.2680	1.4965	1.2732	1.2345	.9881	1.1286

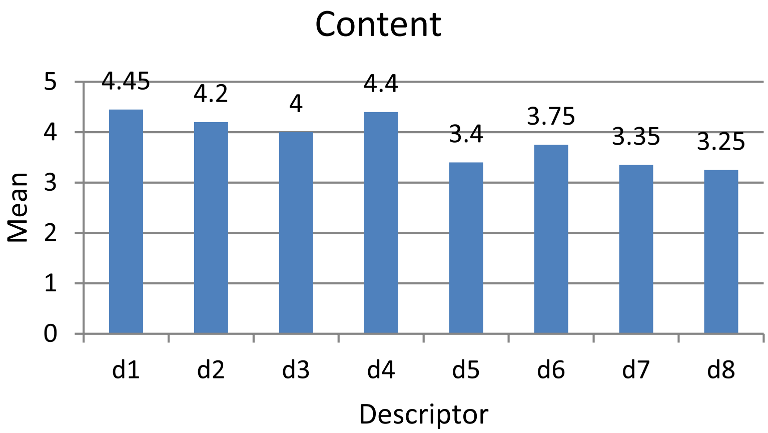

Fig. 2 The degree of importance attached to the descriptors in the content category

Results of the questionnaire on descriptors' importance

The questionnaire posed the question: 'How much importance would you attach to the following descriptors when assessing an undergraduate student's performance?' See Tables 2, 3, and 4 for details.

As shown in Table 2, the order of the descriptors was modified based on their respective mean scores. Consequently, the content sub-scale descriptors were rearranged according to these mean values.

Also, see Figs. 2, 3, and 4 for details.

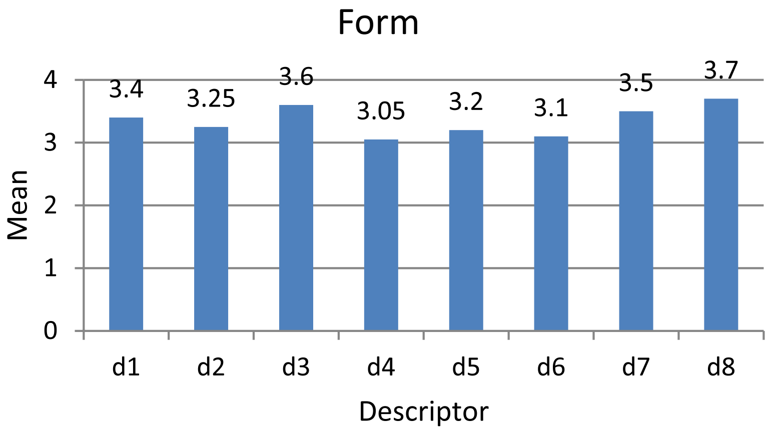

Fig. 3 The degree of importance attached to the descriptors in the form category

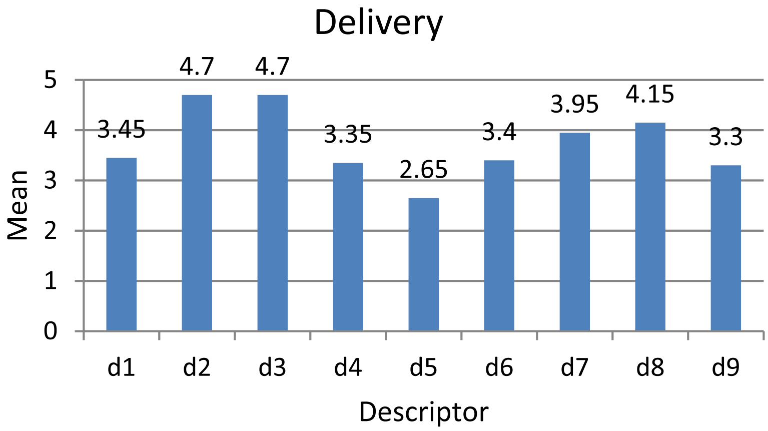

Fig. 4 The degree of importance attached to the descriptors in the delivery category. The descriptor order in the three sub-scales was rearranged based on the above results

Creating the rubric

Based on their importance, the refined descriptors were formatted into a rubric layout. Following Lee's (2015) approach, the 25 remaining descriptors across three criteria categories were integrated into a model rating instrument's layout developed by Christison and Palmer (2005, cited in Bachman & Palmer, 2010). Notably, two descriptors in the delivery section—'The student shows fluency in text/message delivery' and 'The student shows few pauses, hesitations, fillers, and false starts'—received equal weighting. After consultation with experts, these descriptors were combined to streamline the criteria and improve handling efficiency. This decision was supported by recognizing that pauses, hesitations, fillers, and false starts indicate fluency. Finally, the findings obtained from a qualitative study on CI assessment in the Iranian academic context discussed by Shafiei (2021) and the questionnaire results on assessment criteria informed the weighting assigned to each criterion.

Determining the level of effectiveness

Establishing differentiated performance levels supports tailored instructional interventions, enriching CI pedagogy. Although J. Mueller asserted that "there is no set formula for the number of rubric levels, it is commonly recommended to use between three to five scale levels" (personal communication, January 16, 2019). Stevens and Levi (2005) advise limiting rubrics to a maximum of five scales and six to seven dimensions. Oakleaf (2009) notes that an even number of levels (typically 4) is preferable for enforcing evaluative decisions, whereas an odd number (usually 3 or 5) allows for a middle ground. Schreiber et al. (2012) state that performance levels on rating scales can be numeric (scores from 1 to 5), descriptive (e.g., good, fair, low), indicative of behavior frequency (e.g., often, sometimes, rarely), or aligned with another criterion like a grade. In numeric scales, while one is generally the lowest number, zero may be included if appropriate, such as when some students might not include an element.

Preferring a middle ground, the researcher selected an odd number of levels and combined numeric with descriptive levels. Interviews on CI teaching in the Iranian academic context (Shafiei, 2021) revealed that 4 out of 10 interviewees did not use CI techniques, suggesting potential deficits in CI performance and student ability. This finding led to the inclusion of a zero level in the rubric. Sreedharan (2013) notes that a zero level allows evaluators greater scoring flexibility. Therefore, five performance levels with corresponding scores and descriptors were chosen, ensuring detailed feedback for assessors and students. These levels were organized from highest to actual lack of ability.

Clarifying the qualifiers

To minimize divergences in perceptions and inferences, the researcher selected distinct qualifiers for each performance level, ensuring clear differentiation. Fulcher and Davidson (2007) state that judgment methods usually establish cut scores. To enhance transparency, the researcher consulted experts to determine the cut-offs based on the number of descriptors in each criterion. This process resulted in establishing five effectiveness levels with corresponding qualifiers: excellent (7–8 items present), good (5–6 items present), fair (3–4 items present), poor (1–2 items present), and zero (no item present).

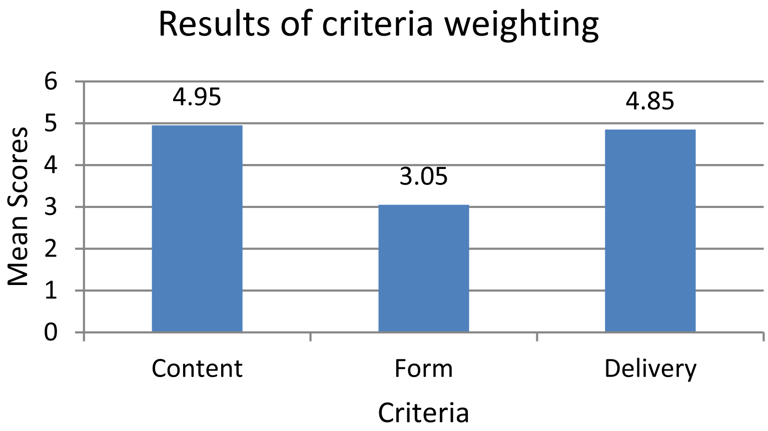

Fig. 5 The total mean score of each criterion

Assigning weight to each criterion

According to the questionnaire results (see Fig. 5), content and delivery are assigned a weight of 2, while form receives a weight of 1. In percentage terms, this translates to 40% for both content and delivery and 20% for form. Thus, content and delivery criteria scores will be doubled based on their level. The questionnaire asked: 'How important do you think each criterion should be in CI assessment at the undergraduate level in Iranian universities?'.

Stating clear objectives for each criterion

Clearly articulating each criterion's objectives fosters student effort and trainer planning. The researcher defined the objectives for each criterion as follows:

- Content: accurate rendition of the source text.
- Form: appropriate target language expression.
- Delivery: successful communicative Interaction.

Qualitative assessment

This rubric section provides a qualitative assessment of the student, offering valuable feedback and diagnostic comments on their performance.

Trying out the rubric: pilot study

The pilot study's role in rubric development highlights the value of empirical testing in educational tool creation (see the Appendix), coupled with colleague review and soliciting student feedback, facilitated the assessment of trainers' understanding of each criterion and their effective use of the rubric. For more details, see Table 5.

The results of inter-rater reliability: pilot phase

See Table 6 for details.

The Pearson correlation coefficient indicated a high correlation ($r = .967$) between the two raters' scores.

Table 5 Students' scores on the CI test

Student no	Rater 1	Rater 2	Average
1	14	13	13.5
2	18	18	18
3	12	10	11
4	15	14	14.5
5	10	10	10
6	15	15	15

Table 6 Inter-rater reliability check-in pilot phase

	Correlations	
	Rater 1	**Rater 2**
Rater 1		
Pearson correlation	1	.967**
Sig. (2-tailed)		.002
N*	6	6
Rater 2		
Pearson correlation	.967**	1
Sig. (2-tailed)	.002	
N*	6	6

*N**Number of students

** $p < .01$

Discussion

The advantage of using rubrics, especially for formative assessment, is enabling students to identify their weaknesses and strengths, thereby reducing objections to final grades. The researcher developed an analytic rubric for interpreting assessment as detailed scoring guide rubrics facilitate the valid assessment of multifaceted performances.

Leveraging the potential of analytic rubrics in assessment settings, the present researcher attempted to develop a tailored analytic rubric for CI within the Iranian academic context. The investigation into the reliability and validity of the proposed rubric, as detailed in the preceding section, revealed its satisfactory reliability and validity. Distinctively, unlike the rubric developed by Lee (2015), which assigned a double weight to 'content' compared to 'delivery' and 'form,' the rubric this study equally prioritized 'content' and 'delivery.' Such a result mirrors the specific needs and priorities of the Iranian academic setting, thereby contributing to a more contextually relevant assessment tool. It also resonates with the findings discussed by Shafiei (2021), who reported significant deficiencies in students' delivery skills as identified by interpreting trainers. Such weighting on the respondents' behalf may mean that Iranian students' delivery aspect must be emphasized more in academic settings to remedy the relevant deficiencies.

Based on socio-cultural and critical approaches to language testing, the context and purpose of assessment critically influence the validity of any rating instrument. Scholars such as Shaw and Weir (2007) and Weigle (2002) advocate developing rating scales tailored to their specific contexts and purposes. Although this study initially aimed to thoroughly apply Lee's (2015) rubric development process, limitations imposed by the context necessitated adopting a different yet feasible data collection method akin to those used in comparable studies.

Assessment is a critical aspect of education for supporting teaching and learning processes. Gipps (1994) highlighted the multifaceted goals of assessment in educational courses, including its role in supporting instruction, providing feedback on learners, teachers, and schools, serving as a selection and certification device, and functioning as an accountability measure. Moreover, effective assessment methods can significantly enhance curriculum and teaching methodologies. Therefore, this study is hoped to encourage further endeavors in assessment practices in CI, establishing a rich area for study and research in interpreting teaching advancement.

Employing sound assessment strategies ensures practitioners in translation and interpreting fields achieve the standards necessary for accurate and effective cross-lingual and cultural communication. This focus on comprehensive assessment acknowledges the complexities of translation and interpreting, underscoring the need for specialized evaluative criteria tailored to the specific demands of each language activity in learning contexts as stepping stones for entering professional spheres. The absence of robust, standardized assessment frameworks hinders the professional growth of aspiring interpreters and affects the quality of interpreting services offered. Consequently, academia must invest in developing assessment strategies, a gap this research aimed to fill by proposing an analytic rubric for CI assessment.

Although the rubric has shown to be reliable and valid, it is recommended that future studies focus on its refinement, including adjustments to descriptors, weightings, and validation procedures, while exploring other prevalent methods in the validation process. The enumeration of limitations related to the development of the rubric, which will be discussed in the subsequent section, underscores the need for methodological enhancements and revisions. Despite its imperfections, this pioneering research underscores the importance of further exploration into teaching and assessment practices in CI within academic settings, aiming to enrich the discourse for researchers and educators alike.

Conclusion

This study notably contributes to the field of interpreting performance assessment, particularly in the context of descriptor-based scales, an area that Han (2017) has identified as "a significant gap in the literature" (p. 198). Aligning with the studies (e.g., Lee, 2008; Tiselius, 2009; Wang et al., 2015) that advanced empirical research on rating scales, this study extends the discourse by explicitly addressing the challenge of ensuring measurement validity in the presence of impressionistic, intuition-based raters. Han's (2016) analysis of 447 interpreting research papers underscores the underrepresentation of rater reliability in existing literature, a critical aspect this study seeks to address by advocating for more comprehensive reporting in rater-mediated measurement research. Central to this study's findings is developing and validating an empirically based rubric, marking a significant stride in offering trainers a more objective and systematic assessment tool. The rubric introduced represents an effort to transition from subjective evaluations to a more systematic approach.

While the findings of this study may contribute insights to the field of interpreting performance assessment, it is important to acknowledge several limitations that may have influenced the results and their interpretation. (1) Limited scope of participant pool: this study's focus on Iran's interpreting research field, a relatively nascent area in academic discourse, inherently limited the participant pool. The burgeoning nature of interpreting studies within the Iranian context has resulted in a relatively small community of experts, as reported previously in a study by Shafie and Barati (2015). Consequently, the limited number of available and willing participants from this specialized field may have affected the diversity and representativeness of the study's findings, potentially impacting their generalizability to broader contexts. (2) Potential subjectivity in qualitative analysis: despite efforts to remain objective, personal biases might inadvertently influence the analysis. (3) Constraints in rubric design: the rubric developed, while empirically based, might have limitations in its design or application. It may not fully capture all the nuances of interpreter performance or could be limited in addressing diverse interpreting scenarios. Looking forward, the implications of this research extend beyond its current scope, paving the way for future inquiries. It is essential for subsequent research to revisit and refine the proposed rubric and expand the participant base to enhance its applicability and robustness. Furthermore, the interplay between research and practical application must continue to be a focal point, with future studies potentially exploring experimental applications of this rubric in self-assessment contexts within CI courses. Such endeavors will further elucidate the nuances of rater reliability and its pivotal role in interpreting assessment methodologies.

Appendix

Analytic rubric developed for CI assessment in B.A. English Translation

Student ID:	**Performance levels**				
	4	**3**	**2**	**1**	**0**
Performance criteria* ***Corresponding descriptors arranged based on the order of importance**	**Excellent 7–8 items**	**Good 5–6 items**	**Fair 3–4 items**	**Poor 1–2 items**	**Zero No item**
A: Content Main Objective: Accurate rendition of the source text 1. The student accurately renders the source text message (s) into the target text. ○ 2. The student renders the source text message(s) thoroughly. ○ 3. The student produces a coherent and meaningful rendering of the source text message. ○ 4. The student avoids any unjustified changes in the information content. ○ 5. The student renders figures, dates, and proper names accurately. ○ 6. The student avoids unnecessary additions. ○ 7. The student avoids unnecessary omissions. ○ 8. The student avoids unnecessary substitutions. ○	Student's level in the content: Rater's comment on content:				

	Performance levels				
	4	**3**	**2**	**1**	**0**
Performance criteria* ***Corresponding descriptors arranged based on the order of importance**	**Excellent 7–8 items**	**Good 5–6 items**	**Fair 3–4 items**	**Poor 1–2 items**	**Zero No item**
B: Form Main objective: Appropriate target language expression 1. The student avoids literal translation. ○ 2. The student uses the same grammatical person as the speaker. ○ 3. The student uses an appropriate register. ○ 4. The student produces a grammatically correct rendering. ○ 5. The student reproduces the logical cohesion of the source text in the target text. ○ 6. The student avoids terminological errors. ○ 7. The student avoids lexical errors. ○ 8. The student avoids phonological deviations. ○	Student's level in the form: Rater's comment on the form:				

	Performance levels				
	4	**3**	**2**	**1**	**0**
Performance criteria* ***Corresponding descriptors arranged based on the order of importance**	**Excellent 7–8 items**	**Good 5–6 items**	**Fair 3–4 items**	**Poor 1–2 items**	**Zero No item**
C: Delivery/presentation Main objective: Successful communicative interaction 1. The student shows fluency in text/message delivery (few pauses, hesitations, fillers, false starts). ○ 2. The student gives an impression of confidence (through eye contact and appropriate gestures and posture). ○ 3. The student uses a clear (no mumbled) articulation. ○ 4. The student renders the source text/message at an appropriate delivery rate. ○ 5. The student has precise pronunciation. ○ 6. The student avoids backtracking and self-correction. ○ 7. The student delivers an interpretation with the correct intonation. ○ 8. The student has an appropriate voice volume. ○	Student's level in the delivery: Rater's comment on the delivery:				

Student's final score

Content Level × 2 + Form Level + Delivery Level × 2 =

Abbreviations

B.A	Bachelor of Arts
CI	Consecutive Interpreting
TL	Target Language
Ph.D.	Doctor of Philosophy
AIIC	International Association of Conference Interpreters

Acknowledgements

Not applicable.

Author's contributions

The author read and approved the final manuscript.

Author's information

Shilan Shafiei, a PhD holder in Translation Studies, is an assistant professor of translation studies in the Department of English Language and Literature, Faculty of Letters and Humanities, Shahid Chamran University of Ahvaz, Ahvaz, Iran. Her research interests include interpreting, translation and interpreting teaching and assessment, inter-semiotic translation, media accessibility, translation and ideology, and translation and literature.

Funding

Not applicable.

Availability of data and materials

Data are available upon reasonable request.

Declarations

Competing interests

The author declares no competing interests.

Received: 17 November 2023 Accepted: 24 February 2024
Published online: 12 April 2024

References

Anderson, L. (1994). Simultaneous interpretation: contextual and translation aspects. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the gap: empirical research in simultaneous interpretation* (pp. 101–120). John Benjamins.

Angelelli, C. V. (2009). Using a rubric to assess translation ability: defining the construct. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice* (pp. 13–47). John Benjamins. https://doi.org/10.1075/ata.xiv.03ang

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: developing language assessments and justifying their use in the real world*. Oxford University Press.

Bartłomiejczyk, M. (2007). Interpreting quality as perceived by trainee interpreters: self-evaluation. *The Interpreter and Translator Trainer, 1*(2), 247–267. https://doi.org/10.1080/1750399X.2007.10798760

Bontempo, K., & Hutchinson, B. (2011). Striving for an "A" grade: A case study of performance management of interpreters. *International Journal of Interpreter Education, 3*, 56–71.

Carroll, J. B. (1966). An experiment in evaluating the quality of translations. *Mechanical Translations and Computational Linguistics, 9*(3 & 4), 55–66. https://aclanthology.org/www.mt-archive.info/MT-1966-Carroll.pdf.

Clifford, A. (2001). Discourse theory and performance-based assessment: two tools for professional interpreting. *Meta, 46*(2), 365–378. https://doi.org/10.7202/002345ar

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge University Press.

DeVellis, R. F. (2003). *Scale development: theory and applications* (2nd ed.). Sage.

Emam, A. (2013). *Applied issues in interpreting*. Shahid Chamran University Press.

Ferdowsi, S. (2014). Moving towards an objective scoring assessment in interpreting. *Iranian EFL Journal, 10*(4), 400–415.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: an advanced resource book*. Routledge.

Gile, D. (2009). *Basic concepts and models for interpreter and translator training* (revised). John Benjamins.

Gipps, C. (1994). Developments in educational assessment: what makes a good test? *Assessment in Education, 1*(3), 283–292. https://doi.org/10.1080/0969594940010304

Han, C. (2016). Reporting practices of rater reliability in interpreting research: a mixed-methods review of 14 journals (2004–2014). *Journal of Research Design and Statistics in Linguistics and Communication Science, 3*(1), 49–75. https://doi.org/10.1558/jrds.29622

Han, C. (2017). Using analytic rating scales to assess English-Chinese bi-directional interpreting: a longitudinal Rasch analysis of scale utility and rater behavior. *Linguistica Antverpiensia, New Series: Themes in Translation Studies, 16*, 196–215. https://doi.org/10.52034/lanstts.v16i0.429

Han, C., & Lu, X. (2021). Interpreting quality assessment re-imagined: the synergy between human and machine scoring. *Interpreting and Society, 1*(1), 70–90. https://doi.org/10.1177/27523810211033670

Hinton, P. R., Brownlow, C., McMurray, I., & Cozens, B. (2004). *SPSS explained*. Routledge.

Huot, B. (1990). Reliability, validity, and holistic scoring: what we know and what we need to know. *College Composition and Communication, 41*, 201–213.

Iriani, T., Anisah, Y. L., Maknun, J., & Dewi, N. I. K. (2023). Analytical rubric development design for objective test assessment. *ACEIVE 2022: Proceedings of the 4th Annual Conference of Engineering and Implementation on Vocational Education, ACEIVE 2022, October 20 2022, Medan, North Sumatra, Indonesia* (p. 327). European Alliance for Innovation.

Knoch, U. (2007). *Diagnostic writing assessment: the development and validation of a rating scale [Thesis]*. The University of Auckland.

Knoch, U. (2009). Diagnostic assessment of writing: a comparison of two rating scales. *Language Testing, 26*(2), 275–304. https://doi.org/10.1177/0265532208101008

Kola, I. M. (2022). Using analytical rubrics to assess technological solutions in the technology classroom. *International Journal of Technology and Design Education, 32*(2), 883–904.

Larson, M. L. (1998). *Meaning-based translation: a guide to cross-language equivalence* (2nd ed.). University Press of America.

Lee, J. (2008). Rating scales for interpreting performance assessment. *The Interpreter and Translator Trainer, 2*(2), 165–184. https://doi.org/10.1080/1750399X.2008.10798772

Lee, S. B. (2015). Developing an analytic scale for assessing undergraduate students' consecutive interpreting performances. *Interpreting, 17*(2), 226–254. https://doi.org/10.1075/intp.17.2.04lee

Li, W. (2022). Scoring rubric reliability and internal validity in rater-mediated EFL writing assessment: insights from many-facet Rasch measurement. *Reading and Writing, 35*(10), 2409–2431. https://doi.org/10.1007/s11145-022-10279-1

Liu, M., & Chiu, Y. H. (2008). Assessing source material difficulty for consecutive interpreting: quantifiable measures and holistic judgment. *Interpreting, 11*(2), 244–266. https://doi.org/10.1075/intp.11.2.07liu

McMillan, J. H. (1997). *Classroom assessment: principles and practice for effective instruction*. Allyn and Bacon.

McNamara, T. (1996). *Measuring second language performance*. Longman.

Moskal, B. M. (2019). Scoring rubrics: what, when, and how? *Practical Assessment, Research & Evaluation*, *7*(3). https://doi.org/10.7275/a5vq-7q66

Nelson, N. W., & Van Meter, A. M. (2007). Measuring written language ability in narrative samples. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 23*(3), 287–309. https://doi.org/10.1080/10573560701277807

Oakleaf, M. (2009). Using rubrics to assess information literacy: an examination of methodology and inter-rater reliability. *Journal of the American Society for Information Science & Technology, 60*(5), 969–983. https://doi.org/10.1002/asi.21030

Pallant, J. (2011). *SPSS survival manual: a step-by-step guide to data analysis using SPSS* (4th ed.). Allen & Unwin.

Pöchhacker, F. (2001). Quality assessment in conference and community interpreting. *Meta, 46*(2), 410–425. https://doi.org/10.7202/003847ar

Pöchhacker, F. (2004). *Introducing interpreting studies*. Routledge.

Pöchhacker, F. (2015). *Routledge encyclopedia of interpreting studies* (1st ed.). Routledge.

Reiss, K. (2000). *Translation criticism – the potentials and limitations: categories and criteria for translation quality assessment (E. F. Rhodes, Trans.)*. St. Jerome Publishing.

Riccardi, A. (2002). Evaluation in interpretation: macro criteria and micro criteria. In E. Hung (Ed.), *Teaching translation and interpreting 4: building bridges* (pp. 115–126). John Benjamins. https://doi.org/10.1075/btl.42.14ric

Russo, M. (1995). Self-evaluation: the awareness of one's difficulties as a training tool for simultaneous interpretation. *The Interpreters' Newsletter, 6*, 75–86.

Saldanha, G., & O'Brien, S. (2013). *Research methodologies in translation studies*. Jerome Publishing.

Sawyer, D. B. (2004). Fundamental aspects of interpreter education: curriculum and assessment. *John Benjamins*. https://doi.org/10.1075/btl.47

Schreiber, L. M., Paul, G. D., & Shibley, L. R. (2012). The development and test of the public speakingcompetence rubric. *Communication Education, 61*(3), 205–233. https://doi.org/10.1080/03634523.2012.670709

Shafiei, S. (2021). *Consecutive interpreting teaching and assessment in the Iranian academic setting [Paper presentation]*. London: TELLSI TS Symposium.

Shafiei, S., & Barati, H. (2015). The state of interpreting studies in Iran: an overview of papers and theses. *Translation Studies, 13*(50), 23–44.

Shafiei, S., Tavakoli, M., & VahidDastjerdi, H. (2017). Delving into the note-taking technique in consecutive interpreting: academic context in focus. *Translation Studies Quarterly, 14*(56), 39–56.

Shafiei, S., Tavakoli, M., & VahidDastjerdi, H. (2019). Employing consecutive interpreting techniques through task-based approach: a case of Iranian learners. *SKASE Journal of Translation and Interpretation, 12*(1), 48–67.

Shaw, D. S., & Weir, J. C. (2007). *Examining writing: research and practice in assessing second language writing*. University of Cambridge ESOL Examinations. Cambridge University Press.

Sreedharan, J. (2013). Not using a zero in evaluation rubrics leads to spurious interpretations. *Gulf Medical Journal, 2*(1), 37–48.

Stevens, D. D., & Levi, A. J. (2005). *Introduction to rubrics: an assessment tool to save grading time, convey effective feedback, and promote student learning*. Stylus Publishing.

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about assessment in higher education: a review. *Assessment & Evaluation in Higher Education, 30*(4), 325–341. https://doi.org/10.1080/02602930500099102

Taras, M. (2005). Assessment- summative and formative-some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466–478. https://doi.org/10.1111/j.1467-8527.2005.00307.x

Tiselius, E. (2009). Revisiting Carroll's scales. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies: a call for dialogue between research and practice* (pp. 95–121). https://doi.org/10.1075/ata.xiv.07tis

Uludag, P., & McDonough, K. (2022). Validating a rubric for assessing integrated writing in an EAP context. *Assessing Writing, 52*, 100609. https://doi.org/10.1016/j.asw.2022.100609

Wadensjö, C. (1998). *Interpreting as Interaction* (1st ed.). Addison Wesley Longman.

Wang, J., Napier, J., Goswell, D., & Carmichael, A. (2015). The design and application of rubrics to assess signed language interpreting performance. *The Interpreter and Translator Trainer, 9*(1), 83–103. https://doi.org/10.1080/1750399X.2015.1009261

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Wojtczak, A. (2002). Medical education terminology. *Medical Teacher, 24*(4), 357–357. https://doi.org/10.1080/01421590220145699

Zwischenberger, C. (2010). Quality criteria in simultaneous interpreting: an international vs. a national view. *Interpreters' Newsletter, 15*, 127–142.

Publisher's Note