A STUDY ON ASSESSMENT CRITERIA FOR ENGLISH – VIETNAMESE CONSECUTIVE INTERPRETING TESTS*

Tran Phuong Linh**, Do Minh Hoang

VNU University of Languages and International Studies, Pham Van Dong, Cau Giay, Hanoi, Vietnam

Received 05 April 2022
Revised 11 June 2022; Accepted 15 July 2022

Abstract: The study investigates the reliability and user feedback about the rubrics to evaluate English – Vietnamese consecutive interpreting tests taken by undergraduates at VNU-ULIS created by Tran and Do (2022). Five VNU-ULIS raters – two experienced raters and three novice ones – independently rated ten different interpreting tests and provided their feedback on the rubrics. The results reveal the newly created rubrics is mostly considered user-friendly and practical application for interpreting evaluation. Overall, inter-rater reliability, which was presented through Cronbach's alpha and the single measure intra-class coefficient, was acceptable. Besides, the value among the novice raters was higher than that between the two experienced ones. The raters' perception of each quality criterion and their rating process may account for the differences in their score decisions. The findings also suggest further improvements in terms of descriptor wording, weightings and rater training.

Keywords: consecutive interpreting tests, assessment criteria, rubrics

1. Introduction

Regarding effectiveness testing of evaluating tools, Lee (2008) pointed out that the central concepts in all assessments are validity and reliability. The validity is "the extent to which we can interpret a test score as an indicator of the ability we want to measure" (Bachman & Palmer, 1996, p. 21, as cited in Lee, 2008). For example, the level of interpreting skills or interpreting quality should be the only thing being assessed in an interpreting test. Reliability, meanwhile, refers to the consistency of measurement of one assessor in various tests or many different assessors in one test. These are "two interlocking concepts" in which test score is nothing without reliability and reliability is needed to facilitate validity (Sawyer, 2004; Moore & Young, 1997). The study is limited to the reliability of the new interpreting quality assessment set for English-Vietnamese consecutive interpreting established in Tran and Do (2022)'s study.

It should be noted that the goal of assessment criteria developed in Tran and Do (2022)'s study was to evaluate undergraduate students' competence to provide a good interpreting performance in their final test of an interpreting course, rather than assessment for real-life interpreting or accreditation. By 'good' performance, the researchers referred to the ability to synthesize information from a complicated type of speech and to deliver accurate interpretation fluently, confidently

* This research has been completed under the sponsorship of the University of Languages and International Studies (ULIS, VNU) under the Project N.21.07

** Corresponding author.
Email address: linhtp@vnu.edu.vn

https://doi.org/10.25073/2525-2445/vnufs.4795

and coherently[1]. In this study, the researchers focus on the effectiveness of the established assessment benchmark to evaluate undergraduate students' English-Vietnamese interpreting performances in the Consecutive Interpreting course's final test at VNU-ULIS.

2. Literature Review

2.1. Consecutive Interpreting

Consecutive interpreting is one of the three modes of interpreting that make up conference interpreting (AIIC, 1982).

> It involves listening to what someone has to say and then, when they have finished, reproducing the same message in another language. The speech may be anything between a minute and 20 minutes in length and the interpreter will rely on a combination of notes, memory and general knowledge to recreate their version of the original. This form of consecutive is sometimes called "long consecutive" to distinguish it from "short consecutive", which usually involves a speaker stopping after each sentence (or a couple of sentences) for the interpreter to translate. (Gillies, 2019, p. 1)

CI used to be "the standard for international meetings of every kind" before being overtaken by SI or, to be more exact, the invention of equipment to make SI possible in the 1920s. However, CI has not vanished. In fact, there are numerous situations in which CI is required, such as ceremonial speeches, business trips, guided tours, high-level bi-laterals, negotiations, depositions / court testimony, press conferences, company in-house training courses, etc. CI will be used whenever complicated equipment for SI is unsuitable or impossible in the setting in which the meetings take place. For example, government visits or technical experts' working trips often involve going around many locations to see how things have been done in another country and the interpreter(s) is expected to get out of the "fully equipped conference centre" to accompany the clients. It means the interpreter(s) has to interpret all descriptions and explanations consecutively. Or in another case, CI is also the only option because the simultaneous equipment has broken down.

There are other reasons why CI is chosen rather than the fact that the setting does not allow SI. Firstly, it is widely believed that CI may "achieve greater accuracy of interpretation" (Van Hoof, 1962; Weber, 1989, as cited in Gillies, 2019; Hale, 2007, pp. 27-30; Pienaar & Cornelius, 2015, pp. 199-200). Because of the nature of CI mode, in which the interpretation is produced after the speaker(s) have finished, the CI interpreter may have time (even it is only a few seconds before starting interpreting) to analyze the SL Content and clarify vague information by asking the speaker directly. Therefore, the interpreter is "less likely to fall victim to misguided anticipation", which can happen more easily in SI as the interpretation is delivered nearly at the same time as the SL. Also, note-taking and corrections are totally possible in CI; hence, it is argued that consecutive interpreters can reformulate all the elements of the source speech, including tone and nuances. While SI only requires medium levels of accuracy with a strong focus on Content, CI provides high levels of accuracy including the manner of speaking (Hale, 2007, pp. 27-30). That is why CI is preferred in settings where sensitive issues are discussed or the verbatim records of interpretation are kept for evidence, such as

[1] Cited from Course guide of Advanced Interpreting course at VNU-ULIS.

in legal proceedings. Secondly, CI helps assure the effectiveness of the delivery of the interpretation through the interpreter's interaction with both the speaker(s) and audience. For example, it is totally easy and acceptable for the interpreter to make eye contact with the participants, clarify what has been said as well as manage the discourse turns (Russell & Takeda, 2015, as cited in Gillies, 2019). It is also obvious that CI is chosen because of its convenience and lower costs. Given that SI booths cost more money and a single consecutive interpreter can be responsible for a two-language meeting instead of a team of at least two like in SI, event organizers who want to cut down the expenses will undoubtedly prefer CI to SI (Ouvrard 2013, p. 85, as cited in Gillies, 2019).

In addition to several reasons mentioned above, CI still plays an essential role in interpreting training. In most translation training programmes including ones at VNU-ULIS, CI training is considered the first and foremost element which offers trainees not only fundamental knowledge and skills relating to interpreting but also the foundation for further study on SI. Consequently, the study on this interpreting mode is necessary, especially when the number of papers about CI in the educational context is quite limited.

2.2. Methods for Interpreting Assessment in Educational Context

There are various methods to assess an interpreting performance, including scoring impression, error counts, checklist, analytic scales, etc. (Lee, 2015). Each method has its own advantages and can be used to serve particular assessing purposes. For example, checklist has been widely used for "recording observations", especially in peer assessment and self-assessment (Brown & Abeywickrama, 2010). The way to conduct the assessment and its purposes may be varied, but, in general, two main approaches to evaluate an interpreting performance in the educational context, namely the holistic marking method and the analytic marking method.

Holistic marking method

The holistic method has gained popularity in both training and the industry since the 20th century (Beeby, 2000, p. 185; Valero-Garcés, Vigier, & Phelan, 2014, as cited in Hale et al., 2012, p. 31). In this marking method, the quality of an interpreting performance will be assessed based on the assessor's "overall impression", "against a predetermined total". That is also why this method is often considered "intuitive' and "impressionistic" (Bontempo & Hutchinson, 2011; Lee, 2009). One obvious advantage of the holistic marking method is that it is fairly easy to understand and usually used for large scale assessment in which no "specific reference to performance features" or just decision about pass or fail is required. (Hunter et al., 1996, as cited in Lee, 2015). However, it is true that holistic marking method would not be a wise choice for inexperienced raters or the ones who are not skillfully trained in it. Green and Hawkey (2010) pointed out that there were radical differences among raters rating the same interpreting performance from holistic perspective because the raters had distinct views on how to interpret and what a good performance should be. As a result, different raters may give different weights to a particular feature of the performance.

Analytic marking method

The nature of this method was pointed out by Mariana et al. (2015, p. 155) in her definition relating the translation testing:

> [The analytic method] is a way to assess the quality of a translation by looking at segments of the text [...] and **awarding or deducting** points

to the overall score of the translation based on whether each text unit meets certain criteria.

In interpreting testing, instead of segments of text, the raters usually look at separate interpreting parts across criteria and finalize the score by using all the various points. So, it can be seen that unlike the holistic method in which the score is generated through overall impression, the analytic one asks the raters to give points based on "pre-specified" criteria. Besides, the definition provided by Mariana et al. (2015) above also mentioned two distinct ways to rate an interpreting performance using the analytical marking method which are (1) error deduction and (2) criterion-referencing (awarding with the use of scales of descriptors). In some translation tests, the combination of these two can be employed to assess the translation quality (Turner et al., 2010, as cited in Hale et al., 2012, p. 53).

Rubrics-based marking method

It is undeniable that the error-focused marking system has become less popular recently. Instead, the scholars' attention tended to switch to criterion-referencing, especially rating scales or rubrics (Angelelli, 2007, 2009; Lee, 2005; Lee, 2008; Jacobson, 2009; Lee, 2015) or both methods. The rating scales or rubrics contain one or some criteria of the skills being assessed and several sets of descriptors of the criteria plotted against levels of achievement. A numeric grade or titles like 'poor', 'good', 'excellent' are assigned to each level. In fact, at an early stage in the marking process the assessor still needs to identify the errors that the assesses made. However, the score of an interpreting performance is awarded as the assessor compares and chooses the level which has the most closely matched descriptors in the rubrics.

One of the strengths of the rubrics-based marking method is that it provides assessors with a bigger and more comprehensive view while evaluating the quality of interpreting performance. In other words, the assessors have to consider "a wide range of factors", like meaning, terminology, style, delivery and the ability of the interpreter, etc., to decide the final score. Because of that, this marking method allows interpreting to be assessed as not only a product but also a process. Secondly, because the rubrics are always attached with a full set of descriptors for each distinguished level, it can be useful for the post-marking stage. For example, trainer(s) can use descriptors, well-designed and written only, to write a detailed and meaningful report of results for their students. It is also easier to justify the result by pointing to the descriptors selected in case there are any disputes from students over the testing results.

However, the rubric-based approach still has its potential flaws. Firstly, to have the rubrics valid, assessing criteria and construct has to be chosen and devised carefully (Angelelli, 2009, p. 22). Secondly, it is important to revise and carefully write the descriptors because any vague terms or expressions can lead to room for differences in what each level means. Besides, because it is obvious that the rubrics-based method is more complicated than the holistic method. An extensive and intensive training and trial marking are required for all raters before officially conducting evaluation.

Notwithstanding these above disadvantages, rating scales or to be more specific, rubrics still showed it usefulness in subjective assessment like language proficiency or interpreting quality performances. The authors firmly believed that the advantages outweigh its disadvantages and use of rubrics is a good method of assessment in this study. Firstly, because assessment means measurement, rubrics with specialized bands for separate criteria will absolutely allow numeric results to be used in interpreting summative

assessment at VNU-ULIS. Secondly, descriptors in rubrics can not only assure the comprehensiveness but also be a more reliable source for raters who do not have much experience in their assessment.

2.3. Consecutive Interpreting Assessment Criteria

This study used the findings in Tran and Do (2022)'s study in which a set of criteria was created for CI test marking purpose in educational context. From 132 criteria in total for interpreting assessment, the researchers analyzed and created the rubrics with three macro criteria which also received considerable agreement among various scholars including content fidelity, target language quality, delivery. This assures there would not be too much pressure on evaluators' memory to memorize all the criteria but still provides adequate explanation for each to produce more correct numeric assessment.

Content fidelity

Content fidelity is most widely used and "invariably deemed essential" in interpretation rating practice. (Liu, Chang, & Wu, 2008, as cited in Liu, 2013; Pöchhacker, 2001). This concept, in fact, has been given different names and expressions by different researchers. In this study, Content fidelity includes accuracy and completeness of information or meaning transferred from the source language to the target language. This definition was taken from the description of the rating scales used in *Chinese and English Translation and Interpretation Competency Examinations* (*ECTICE* exams) for both translation and CI (Liu, 2013).

'Accuracy' was mentioned in previous studies by various terms. Bühler (1986), Gile (1995, 1999), Kurz (2001) claimed sense consistency with original messages is a fundamental aspect of interpreting quality. This concept is close to the core quality criteria "accurate rendition" included in Roberts' (2000) research, Pöchhacker's (2001) model and Lee's (2015) interpreting feedback form. It is also similar to "equivalence" proposed by Riccardi (2002) or "equivalent intended effect" in Pöchhacker's (2001) and includes *intelligibility* and *informativeness* in Carroll (1966)'s study. The quality of accuracy here should "go beyond lexical similarity between the source speech and the interpreted rendition" (Lee, 2008). The interpretation is considered as a "faithful image" (Gile 1991, p. 198) or "exact and faithful reproduction" (Jones, 1998, p. 5) of the original discourse. As it is based on the accurate comprehension of the source speech, the interpreted version may achieve the same effect on the target audience as on the source language audience.

In addition to 'accuracy', 'completeness' also plays a crucial role in interpreting assessment (Kurz, 2001). The completeness of an interpreting performance, at first, is generally defined as the percentage of what is reproduced among all propositional units, as if CI is a variation of typical prose recall (for example, Goldman & Varnhagen, 1986; O'Brien & Myers, 1987). In ecologically sound interpreting research, however, the concept of completeness may be more *trained* and contextualised to assessors. It is not the relative completeness, but the optimal completeness that interpreters should aim to achieve most of the time, particularly in CI. This idea was succinctly captured by Setton (2005, p. 288, as cited in Setton & Dawrant, 2016) in a relevance-theoretical framework:

> Any global measure of quality should therefore include a measure of procedural effectiveness, i.e. of how effectively the interpreter's discourse evokes the relevant context, in addition to the traditional check on whether information explicitly encoded is sufficiently explicitly rendered. Recognising the

> role of inference in communication will lead to a very different assessment of completeness: for example, referents not explicitly reproduced in the output will not be penalised as omissions if they are easily inferable. (Jin, 2017, p. 8)

In the testing system, the level of fidelity may be reflected in the extent to which certain deviations are observed in interpreting performance. These deviations are meaning units, unjustified changes or distortions, omission, additions of the meaning and intention, and logical cohesion. They are easily recognizable and countable; hence the marking can be done more precisely and easy to seek agreement among different raters. Besides, 'coherence' or idea organization is also an important element in meaning transferring. This sub-criterion examines whether the interpreted text is arranged in an orderly and consistent manner and whether the different parts of the oral rendering are well integrated into a whole (Ouyang, 2017).

It is also important to note that the rating units for this criterion cannot be individual sentences as individual sentences can "convey at least the 'core' meaning" in the original sentences (Carroll, 1966, p. 56). Considering the nature of consecutive interpretation and the difficulty to match individual target language sentences with source language sentences, a segment of several sentences that cohesively form an idea is regarded as the rating unit. This is also suitable with the final interpreting test structure at VNU-ULIS, in which the whole consecutive interpreting is divided into four to five segments. Each segment can contain two to three meaning units.

Target language quality or language use

Target language quality or *target language use* is associated with "the quality of the rendition in terms of linguistic 'correctness', naturalness, and contextual appropriateness of language" (Lee, 2008). This criterion is similar to the "adequate target language expression" in Pöchhacker (2001)'s model, "form" in Lee (2015)'s feedback form and "target text features" in Wang et al. (2015). It should be noted that Content fidelity and Target language quality are two independent criteria. While the former focuses on the equivalence and amount of information rendered from the source to target speech, the latter targets specifically at the quality of the target language. A qualified interpretation should achieve a greater level of not only terminology and grammar accuracy but also tone and nuances[2]. It should sound a newly created speech in the target language.

The sub-criteria for target language quality are features of terminology, naturalness or idiomatic target-language expressions, grammatical aspects, and register and style. Ineffective source language interference may be included in this criterion.

Delivery

According to Pöchhacker (2001), the quality of an interpreting performance cannot be "pinned down" to linguistic aspects only. The level of communicative effect and impact on the interaction within particular situational and institutional constraints must be considered during the evaluation process. Sharing the same opinion, Wadensjö (1998, p. 287) said "in practice, there are no absolute and

[2] Consecutive interpreters hear the entire utterance before delivering the interpretation, which means they can take more time to reformulate all the elements of the source speech and it is easier for bilinguals who are present to notice interpreting errors. Therefore, all the elements of language, including linguistic accuracy, tone and nuances, in CI are often assumed and required to be at a higher level than those in SI. (Russel & Takeda, 2014)

unambiguous criteria for defining a mode of interpreting which would be 'good' across the board. Different activity-types with different goal structures, as well as the different concerns, needs, desires and commitments of primary parties, imply various demands on the interpreters."

Unlike the other two criteria, content fidelity and target language quality, the third criterion under the name of "delivery*"* can be assessed without reference to the source language or source speech. Delivery features invoke public speaking or, more broadly speaking, effective communication skills (Lee, 2008). It is similar to "successful communicative interaction" in Pöchhacker (2001)'s model.

The criterion 'delivery' in this study measures a number of different elements. Fluency is the first and foremost one. This element can be shown through the number of hesitations, length of pauses, frequency of fillers and false starts, repetitions and self-corrections. A successful delivery can also be measured through other obvious elements such as articulation, voice, confidence and pace. Articulation is also known as pronunciation, diction or enunciation, in which in a good interpreting performance, all the words have to be pronounced correctly with appropriate stress and intonation, pleasant voice and easy-to-hear volume. Confidence or poise is a recognizable element and affects the effectiveness of every communication situation. As a result, it is also a sub-criterion in this assessment category. Confidence indicates in both speaking manner and how the interpreter respond to errors during their interpreting process. 'Pace', in this research, refers to the interpreter's ability to switch between two languages, time lag and ability to finish the interpretation within time limit. All the tests are audio-taped with pre-set time for interpretation among each segment, the students who took the test were required to start their interpretation from English into Vietnamese as soon as possible and complete it within that amount of time.

While some researchers or existing framework include "accent" as a marking element (ATA Certification Exam & Zwischenberger, 2010*),* it is still a matter of controversy whether it should be considered in the evaluation. The reason is it is really hard to measure how good or acceptable an accent should be. In Vietnam, for example, Vietnamese people in the North may find it difficult to fully understand the speech or interpretation conducted by people with Central or Southern accent as there are three main accents belonging to three main regions of the country. Therefore, the researchers decided to exclude this micro criterion in this research.

All three mentioned macro criteria (content fidelity, target language quality and delivery) were then used to write a rubric for evaluating interpreting final tests at VNU-ULIS. It is a 6-point scale from level '0' to level '5'. All the descriptors start with "action words" to depict the interpreting and linguistic competences which test takers achieve. Here are some verbs used in the rubric: 'convey', 'make', 'organize', 'demonstrate', 'produce', 'show', 'have', 'interpret', 'reflect', 'display', 'lack', 'fail'. The length of rubric does not exceed two A4 pages (in case landscape orientation is applied).

The level of quality was differentiated by the following adjectives: namely, 'skillful, 'good', 'adequate', 'weak' and 'inappropriate'; while the gravity of deviations was differentiated by the quantifiers: 'many', 'some', 'few', 'one', and 'no'. In order to add the specification to distinguish such levels, adverbs are also used: 'entirely', 'frequently', 'mostly', 'very logically', 'logically', 'adequately', 'partially', and 'rarely'. Regarding the weightings, Tran and Do (2022) suggested it should be 50% for content fidelity, 25% for

target language quality and 25% for delivery (for details of the rubrics, see Appendix).

3. Methodology

The aim of this study is to trial the newly developed rubrics through answering the following research questions:

1. *Does the newly developed rubrics produce reliable assessment when assessing the final exam of Consecutive Interpreting course at VNU-ULIS?*
2. *How do the raters comment after using it?*

The two research questions are addressed from both qualitative and quantitative approaches. The quantitative (numeric) data are collected and analyzed through a numerical test, followed by another source of qualitative data from an interview. From both sources of data, the authors can corroborate findings across data sets and thus reduce the impact of potential biases that can exist in a single study.

3.1. Research Participants

Five ULIS interpreter trainers (distinguished by codes R1, R2, R3, R4 and R5 for Rater 1, Rater 2, Rater 3, Rater 4 and Rater 5 respectively) were invited to evaluate ten consecutive interpreting from English into Vietnamese, using the proposed rubric. Among those trainers, R1 and R2, Experienced Raters (ER), have 10 years of experience in interpreting training and assessment, R3, R4 and R5, Novice Raters (NR), have less than four years of experience. The five trainers were chosen for the convenience of the study as they were teaching interpreting at the study period and available to join the research.

In this research, five raters were asked to rate the same ten interpreting performances. These samples were randomly extracted from the audio-taped recordings of the final test of Consecutive Interpreting course by ten different third-year ULIS students specializing in English translation and interpretation. The English-Vietnamese consecutive interpreting test was about five minutes in length (including both source language speech and student interpreting time) and divided into 5 subsections of 30-45 seconds each (source speech length). Besides, the English source was from an authentic video in which was performed by a native British speaker at an average speed of 140 words per minute in his real TV show interview. These recordings were chosen as the data in this study because they were collected from the latest English-Vietnamese interpreting exam at the study time.

Before the testing phase, a 30-minute face-to-face rater training was conducted to make sure that all the raters understood the rubric and how to implement the marking process. However, because of differences in their working timetable and social distance during Covid-19, only one rater could join face-to-face training. The others received a guiding package through email and undertook the marking process at their leisure. The package included the procedure consent form, the newly developed rubric, explanation of the criteria in the rubric, a score sheet, a folder of eleven audio files with source speech and students' interpretations, English source speech transcript and its suggested answer of the interpreting, and a marking sample of a test (see Appendix). The raters were asked to examine the sample carefully and raise questions (if any) before marking the rest of ten interpreting performances. They worked separately without knowing the scores given by other raters. The scores of each marking category had to be written in the score sheet and sent back to the researchers within a week.

After the marking procedure, the raters were all invited to join an interview. The interviews were conducted with three

raters online (through Zoom.us platform) and with two offline (right after they finished their marking). The manner of conducting the interview was semi-structured in which the researchers relied mainly on the given interview question log but still may ask participants to further their answers when necessary. The interview questions were written in English and the interview was expected in English as the participants are all assumed to be proficient users of the language due to their job as English lecturer. In case the participants ask to be interviewed in Vietnamese, the researchers may resort to translating her questions into Vietnamese, and recording all the answers. All the interviews will be recorded using a digital recorder with consent from the interviewees themselves who are informed before the actual interviews begin. During the interviews, the raters were encouraged to share their experience in interpreting training and marking and provide feedback on the newly developed marking rubric.

3.2. Data Analysis

Due to the scope of the study, this research focuses on the reliability of the rating only. Generally, reliability of assessment consists of two types: the consistency among different raters on the same test(s) or inter-rater consistency and the consistency within the same rater across various tests or intra-rater reliability (Bachman & Palmer, 1990, p. 222, as cited in Lee, 2008). However, because testing and retesting of the same test takers and double marking were not undertaken in this study, intra-rater reliability was not examined in this study. Inter-rater reliability is the **level of agreement between raters or judges**. If everyone agrees, IRR is 1 (or 100%) and if everyone disagrees, IRR is 0 (0%). Several methods exist for calculating IRR, from the simple (e.g. percent agreement[3]) to the more complex (e.g. Cohen's Kappa or Fleiss's Kappa)[4].

Cronbach's alpha and intraclass correlation coefficients (ICC) were chosen to measure inter-rater reliability.

Cronbach's alpha

Cronbach's alpha, α (or *coefficient alpha*), developed by Lee Cronbach in 1951, measures reliability, or internal reliability[5].

The formula for Cronbach's alpha is:

$$\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N-1) \cdot \bar{c}}$$

Where:

N = the number of items.

$\bar{c}$ = average covariance between item-pairs.

$\bar{v}$ = average variance.

In this study, Cronbach's alpha was computed using SPSS by following these steps:

Step 1: Click "Analyze," then click "Scale" and then click "Reliability Analysis."

Step 2: Transfer your variables (q1 to q5) into "Items,". The model default should be set as "Alpha."

Step 3: Click "Statistics" in the dialog box.

[3] Percent agreement is number of agreement scores / total scores. However, chance agreement due to raters guessing is always a possibility - in the same way that a chance "correct" answer is possible on a multiple choice test.

[4] Cohen's kappa statistic measures interrater reliability (sometimes called interobserver agreement). Interrater reliability, or precision, happens when your data raters (or collectors) give the same score to the same data item. In addition, Cohen's Kappa has the assumption that the raters are deliberately chosen. If your raters are chosen at random from a population of raters, use Fleiss' kappa instead.

[5] Internal consistency reliability is a way to gauge how well a test or survey is actually measuring what you want it to measure.

Step 4: Select "Item," "Scale," and "Scale if item deleted" in the box description. Choose "Correlation" in the inter-item box.

Step 5: Click "Continue" and then click "OK".

Regarding the results of Cronbach's alpha, the table below suggests a widely accepted interpretation of the results.

Cronbach's alpha	Internal consistency
α >= 0.9	Excellent
0.9 > α >= 0.8	Good
0.8 > α >= 0.7	Acceptable
0.7 > α >= 0.6	Questionable
0.6 > α >= 0.5	Poor
0.5 > α	Unacceptable

Like any inter-rater reliability, a score of more than 0.7 is generally considered acceptable (Shohamy, 1985, p. 70).

Intraclass correlation coefficients

Intraclass correlation measures the reliability of ratings or measurements for cluster data that has been collected as groups or sorted into groups. In other words, the ICC is used to measure a wide variety of numerical data from clusters or groups, including:

- How closely relatives resemble each other with regard to a certain characteristic or traits.
- Reproducibility of numerical measurements made by different people measuring the same thing.

This study takes advantage of the second use. The ICC was also calculated in SPSS with "Two-Way Random" as there were five different raters rating ten tests. The confidence interval is 95%.

Like most correlation coefficients, the ICC ranges from 0 to 1. A high Intraclass Correlation Coefficient (ICC) close to 1 indicates high similarity between values from the same group. A low ICC close to zero means that values from the same group are *not* similar.

In addition to ICC, mean and standard deviation of ratings were also calculated using SPSS.

4. Findings and Discussions

4.1. Reliability of Rating

Cronbach's alpha was high across the board. All the values were above 0.7, which means that the inter-rater reliability is acceptable and good. In Table 1, the criterion Content fidelity constantly received the highest numbers in the combined group of all five raters, as well as within the other two smaller groups of experienced raters (ER) and novice raters (NR); the numbers were 0.936, 0.862 and 0.874 respectively. This indicated inter-rater consistency in ranking this target criterion in spite of the lack of direct or face-to-face rater training. Cronbach's alpha for TL Quality, meanwhile, is the lowest. The value for this criterion was good at 0.904 in the combined group and 0.826 in NR group, but it was only considered acceptable in ER group at 0.717. Surprisingly, inter-rater reliability was higher in the NR group than in the ER group. While the gap between Cronbach's alpha for Content fidelity in the NR group and the ER group was quite small at 0.012, the figure for TL Quality was 0.109, while that for Delivery was 0.16.

Table 1

Inter-Rater Reliability

Cronbach's alpha	Content fidelity	TL quality	Delivery
Combined group	0.936	0.904	0.917
ER	0.862	0.717	0.745
NR	0.874	0.826	0.905

Table 2

Intra-Class Correlation Coefficient

Intraclass correlation coefficient	Content fidelity	TL quality	Delivery
Combined group	0.745	0.652	0.687
ER	0.757	0.559	0.594
NR	0.697	0.613	0.76

In order to check how reliable a rater can be, the single measure intra-class correlation coefficients (with 95% confidence interval) were computed. The results reflected a similar trend with the ones shown in data about inter-rater reliability (see Table 2). In the combined group, the value for Content fidelity was still the highest at 0.745, which was 0.058 and 0.093 higher than that for Deliver and TL Quality respectively. The novice raters seemed to be more internally consistent in the rating process than the experienced group. The correlation coefficients for *TL quality* within the NR group was 0.697, but it was only 0.757 in ER group. It can be seen that the correlation coefficients in the ER group for the other two criteria, TL Quality and Delivery, were under acceptable level at 0.559 and 0.594 respectively. However, the intra-class correlation coefficient in the ER group was higher than the NR group, which was 0.757 for the ER group and 0.697 for the NR group.

Table 3

Mean and Standard Deviation of Ratings by ER Group and NR Group

	ER		NR	
	Mean	Std. deviation	Mean	Std. deviation
S1_CF	1.30	0.00	1.47	0.29
S2_CF	2.40	0.14	2.77	0.25
S3_CF	3.55	0.35	3.47	0.29
S4_CF	2.80	0.71	2.37	0.12
S5_CF	2.15	0.92	2.43	0.40
S6_CF	1.25	0.35	1.13	0.29
S7_CF	1.90	0.14	2.27	0.25
S8_CF	1.50	0.71	1.53	0.25
S9_CF	4.05	0.35	3.37	0.12
S10_CF	1.40	0.14	1.27	0.25
S1_TLQ	2.15	0.21	1.83	0.29
S2_TLQ	2.65	0.50	2.70	0.53
S3_TLQ	3.9	0.14	3.43	0.12
S4_TLQ	3.05	0.35	2.70	0.17
S5_TLQ	2.65	0.21	2.43	0.40

S6_TLQ	2.15	0.21	1.53	0.25
S7_TLQ	2.65	0.50	2.20	0.36
S8_TLQ	2.5	0.71	1.77	0.25
S9_TLQ	3.65	0.50	3.27	0.25
S10_TLQ	2.25	0.35	1.53	0.25
S1_D	2.15	0.50	1.93	0.51
S2_D	3.30	0.00	3.17	0.29
S3_D	4.40	0.14	4.03	0.46
S4_D	3.55	0.35	3.03	0.64
S5_D	3.15	0.21	2.93	0.60
S6_D	2.90	0.14	1.60	0.53
S7_D	3.50	0.71	3.00	0.50
S8_D	3.15	0.21	2.43	0.12
S9_D	4.50	0.71	4.43	0.81
S10_D	2.75	0.35	2.03	0.25

(S: Student, CF: Content fidelity, TLQ: Target language quality, D: Delivery)

It can be seen from Table 3 that ERs tended to give higher scores than NRs by a maximum 0.72 (see Student 10's mean scores for criteria Target language quality and Delivery). Looking at the standard deviations for Content fidelity and Target language quality, the variations in ratings were mostly higher in the ER group, indicating that ERs were less reliable in this study.

In short, inter-rater reliability which was examined through two parameters, namely Cronbach's alpha and the single measure intra-class coefficients, was at acceptable level for three assessment categories. The criterion Content fidelity achieved the highest values in both groups, followed by Delivery and TL Quality. The findings also revealed that the ERs were more generous assessors than the NRs, but not as reliable as the NRs in this study.

4.2. Raters' Feedback on the Rubric

Through the interview, five raters participating in this research described their marking process using the newly developed rubrics, gave comments about it and suggested significant changes. Three outstanding features of the rubric have been listed and explained as below.

Firstly, all three macro-criteria in the rubrics were perceived to be adequate and comprehensive. The newly developed rubric was commented to "include the most important criteria in interpreting quality assessment" (Rater 1). All raters showed their agreement with the proposed criteria because they are also the ones they used to mark their students' interpreting performances before this research "but perhaps under different terms" (Rater 3).

Secondly, most participants found the rubric useful and fairly user-friendly. According to Rater 1 and Rater 2, the descriptors were written "quite detailed with highlighted keywords to emphasize the differences among different bands". This is better than "vague guidelines" which is "a usual challenge" for inexperienced raters

(Rater 5). It is considered as "a huge advantage" (Rater 4) because after thorough study on the newly rubrics and sample marking, the rater could "produce immediate score" when hearing the students' interpreting and looking at the band descriptors. Some raters thought the rubric is "helpful" as they can "quantify" the interpreting quality, hence it seems to be "reliable" (Rater 1, Rater 2, and Rater 4). Besides, Rater 2 commented that the rubric was still "fairly user-friendly" though it took her a lot of time to understand and differentiate five bands in each assessment criteria.

Thirdly, despite different techniques each rater chose to use, the study showed that marking ULIS interpreting tests using the rubric required a reasonable amount of time. The raters responded they needed to hear the full recordings once and occasional incomplete segments to mark the tests. Rater 4 and Rater 5, who conducted the marking process offline, listened to all ten recordings once only and finished the evaluation within 55 minutes. Rater 2 reported that she decided to hear the first two tests twice "to get familiar with the evaluation using the new rubric "and once for the other recordings. Rater 1 and Rater 5 maintained hearing the audios the second time for a few segments in particular test with different strategies. While Rater 1 paused after each segment in all the tests, Rater 5 listened to full recordings and listened again to some parts to produce scores 'carefully'. No matter how many times the Raters had to listen, it took them about one hour to mark all ten sample tests in this study. This is good time management as ULIS interpreting assessors often have to evaluate up to hundreds interpreting performances including both mid-term and end-of-course tests each semester.

When it comes to details of the rubric, the number of bands was favored by all the raters. Besides, the current descriptors were mainly approved, but one novice rater still preferred more specific descriptions and one experienced rater thought the descriptors were too long. All the raters agreed that the criterion Delivery was the easiest category to evaluate because they had no difficulty in understanding and differentiating five bands in this criterion. As a result, they could make a very quick and precise decision relating to the score of this criterion after only listening once. That the correlation coefficients of Delivery, which is presented in the previous part, was quite high means level of agreement among different raters in this criterion was high. The other two macro-criteria, however, were not that easy for the raters. The main reason lied in the word choice for particular bands and it seemed to be more challenging for the novice raters compared to the experienced ones.

> It is easy to give students band 3 and above, but it is much more difficult to decide between band 1 and 2. When I read the descriptors for Content fidelity, I understood that if students made some minor or one major errors in accuracy and completeness, they will get band 3. Band 2 will be applied if there are some serious errors and band 1 means many serious errors are made. So, what if the student made some major errors and some minor errors at the same time, which band should I give them? I was really confused by this case. (Rater 5)

It can be seen that due to lack of experience in interpreting training and testing, novice raters often had difficulties in distinguishing between minor and major errors and they often strictly based on quantifiers like 'one', 'some' or 'many' to choose suitable bands when marking the interpreting tests. This, however, is less challenging for experienced raters.

> I supposed Content fidelity is the easiest marking category. If it is a perfect interpreting performance in terms of accuracy and completeness, which I could not catch any errors, it is absolutely band 5. In band 4, there are a few but still acceptable mistakes. Band 3 has more errors than band 4… (Rater 1)

Some words or phrases in the rubric are also considered vague and require more thorough teacher training to assure the scoring results among different raters. Rater 2 was confused between '*very logically*' and '*logically*' in idea organization in Content fidelity while Rater 4 and 5 required more examples to understand differences between '*skillful*' and '*good*' use of vocabulary in TL Quality or the meaning of '*idiomatic*'.

At least three out of five raters suggested removing the criterion about grammar from the rubric. Their reason is as Vietnamese is the target language, "the naturalness itself is more important than grammar in Vietnamese' (Rater 2, Rater 3 and Rater 5); plus, it is hard to judge whether Vietnamese grammar was correct or not" (Rater 5).

There was general agreement on weighting among different raters in which all the raters agreed that the criterion Content fidelity should be given the biggest weight (at least 40%). There was greater consensus on weightings among ERs than NRs. Both ERs suggested the ideal weightings be 50% for Content fidelity, 30% for TL Quality and 20% for Delivery. Only one rater thought TL Quality should be given smaller weight than Delivery.

> It's English-Vietnamese interpreting and it occurs to me that Vietnamese is not so highly demanded in terms of grammar and collocations like English. More importantly, I believe that paralinguistic elements in delivery category have more impact on making impression and contributions to a good interpreting performance. (Rater 3)

Table 4

Rater's Views on Weights

	Content fidelity	**TL quality**	**Delivery**
Rater 1	50%	30%	20%
Rater 2	50%	30%	20%
Rater 3	50%	20%	35%
Rater 4	40%	30%	30%
Rater 5	45%	30%	25%

There is one comment on the length and the language of the rubric. It was thought to be better for rater if the rubric was written within one page and in Vietnamese. The shorter the rubric is, the better and faster raters can learn and memorize. Other suggestions were made relating to test design and marking guidelines. The current test includes 5 small segments lasting from 30 to 45 seconds, which was considered "*too short to assess how good the language use was*" (Rater 2). Moreover, through reflection of marking techniques, all raters gave scores for each individual segment and calculated the average to finalize the score for each criterion. This involves a lot of numbers and calculation; therefore, it can be time-consuming and easy to miscalculate. A detailed marking guideline is a must to facilitate the assessment. Acknowledged that some raters, especially novice ones, may strictly follow the marking guidelines, the suggested answer should include more than one way to interpret and point out which answer should not be accepted. A detailed but flexible enough answer can save rater a lot of time and increase inter-rater reliability.

The two raters who had just completed interpreting training, stated that the design of the rating sheet (see Appendix) along with the rubric would help them identify the test takers' good points and

mistakes during their interpreting. These comments suggest that the use of well-defined and well-written rubric may also be useful for formative assessment or in-class activity when trainers have to provide feedback for their trainees.

4.3. Discussion

The study assessed the reliability and users' feedback about the rubrics to assess interpreting tests at VNU-ULIS created by Tran and Do (2022). It should be noted that a marking rubric is a popular tool in evaluation and all three criteria in the rubrics including Content fidelity, Target language quality and Delivery was consistent with the findings in the studies conducted by Lee (2008), Lee (2015) and Wang (2015).

Although rubrics is a reasonable choice, weighting was a controversial area which requires more investigation. That the proposed formula of weightings in which TOTAL = 50% Content fidelity + 25% TL Quality + 25% Delivery was not completely agreed among interviewed raters may be attributed to various personal concepts about the importance of each criterion among raters and special features of the Vietnamese language. From the interviewees' opinion, the authors proposed another formula of 50% + 20% + 30% for Content fidelity, TL Quality and Delivery respectively. This calculation is also suitable when considering characteristics of the target language which grammar does not play an essential role in Vietnamese.

Besides, the word choice during the writing level descriptors stage was the main cause of confusion among interviewed raters. These qualifiers and adjectives like 'very', 'some', 'a few', 'major', 'minor', 'good', 'skillful' should be considered carefully before being added in the descriptors.

In the second phase, the testing of newly developed rubrics, several findings are highlighted as below. The first key finding is ERs turned out to be more generous markers than NRs, particularly Rater 1 with 10-year experience and Rater 5 who has just spent only 14 months in translation training and one semester in interpreting training. This may have been attributed to different perceptions of quality held by ERs and NRs, or different levels of expectation between these two groups of raters. Through the interview, both ERs agreed that experienced interpreting trainers may be 'less demanding' or 'less strict' than those who worked in the industry only. It is mostly because they were aware of the educational context, in which several factors such as expected learning outcomes of the course, test design and students' competence needed to be considered. The NRs, on the other hand, strictly followed the descriptors and to some extent, lack of knowledge about quality from client or users' perspective, can show high expectation about how well test takers must perform in order to reach the top levels on the scale.

Another key finding is that the inter-rater reliability between experienced raters (Rater 1 and Rate 2) was lower than among novice raters (Rater 3, 4 and 5). This finding, which is surprisingly inconsistent with the previous results in Lee (2008)'s study, may be explained by the following factors. It is true that rating can be influenced by raters' backgrounds, experience and expectations, or their different interpretations of scales, standards of severity and reactions to elements not relevant to scales (McNamara, 1996; Lee, 2008; Wang et al., 2015). In this study, with 10 years in interpreting training and testing, both experienced raters have evaluated a wide range of interpreting performances; therefore, they may have established their own personal standards in interpreting quality. Additionally, as both ERs did not attend face-to-face training, there was no chance for discussion and seeking agreement between them as well as

nothing to assure that they had investigated and strictly followed the guidelines to use the newly developed rubric without any different personal assumptions.

Thirdly, content fidelity received the highest inter-rater reliability among three assessment criteria in the newly developed rubric. The reason is fidelity or accuracy or any names it may take is the universal and the most important element in interpreting quality, which was mentioned in all research in this field. The rater's perception of the criterion may be similar; consequently, there would be a high percentage of agreement in each band score. Besides, the fact that the criterion Target language quality has the lowest inter-rater reliability and interview data about weightings suggests that there should be a change in the total weights. Like what has been discussed in the previous session, the authors proposed the lowest proportion for the criterion Target language quality, about 20%. Obviously, the effectiveness of this change still awaits close investigation which is beyond the scope of this study.

Fourthly, at least three out of five raters answered in the interview that they found it difficult to identify the difference between band 2 and band 3 for criterion Content fidelity while one rater claimed she was confused between band 4 and band 5 for criterion TL Quality. This confusion originated from the choice of quantifiers and expressions in the descriptors. Obviously, a review and modification are needed to rewrite all the descriptors, especially those highlighted with confusion and difficulties for raters to understand. Nonetheless, it can be a good idea to decrease the number of levels in the rubric to four levels instead of five like the current one. A four-point scale will not only reduce level overlapping among two successive levels, hence it is easier for the raters to award the appropriate level for performance.

It is noted that rater training plays a critical role in the marking process. A thorough rater training is to make rater feel more assured of the whole marking process and to narrow down the disparities in ratings. A group face-to-face training before the actual evaluation is compulsory to ensure optimal inter-rater reliability and intra-rater reliability. Samples with sufficient varieties for each level on a rating scale or different translated options should be provided to elicit the differences among different bands. It is also important to give raters adequate time for rating practice and discussion can also help raters achieve ease and consistency during actual rating. At the same time, the grading protocols, including suggested timeframes, the marking process and scoring techniques should be agreed upon within the group of assessors.

During the whole marking process, the raters are encouraged to compare different students' performances and scores to maintain their self-consistency and make marking adjustments to scores if necessary. Ideally, all raters should have been gathered at the end of the assessment process to discuss: why they gave highly discrepant scores to some tests, their views on the assessment rubrics and the rating process, and their suggestions for ways to achieve satisfactory inter-rater reliability.

5. Conclusion

The study has made a meaningful contribution to addressing the complexity of interpreting performance assessment, by showing the application of rubrics in assessing English-Vietnamese consecutive interpreting quality in an educational context like summative assessment at VNU-ULIS. It can be concluded from the study that the newly developed rubric in this research might work effectively in multiple interpreting performance assessments, particularly as a means to enhance rating

consistency. By using one standardized rubric with detailed descriptors, summative assessment in interpreting courses as well as general interpreting evaluation would be more correct and consistent among different raters over time. Another contribution made in this study is that novice raters can be more reliable than experienced ones as long as they all have background in interpreter training and a careful rater training is provided.

However, this study has inherent limitations that should be taken into account. A major limitation is that all conclusions regarding inter-rater reliability need to be qualified in light of the small number of raters (N = 5) and the small number of tests evaluated by all five raters (N = 10). A larger sample, more than ten raters and or more than 50 tests, for example, would have enabled the researchers to run a more reliable SPSS analysis.

The findings also indicate that a great deal of additional research remains to be done to validate the rating scales. Follow-up research is also needed in order to implement this rating rubric in a wider context, for example with bigger samples or in a different context or with different groups of raters such as interpreter educators, interpreting practitioners, users. Extensive feedback from rating scale users would be helpful not only in fine-tuning the scale, but also in designing a modified version for a different assessee group and/or a different mode of interpreting. Second, the issue of establishing relative weighting for assessment categories should receive further attention. How relative weighting can be applied to assessment should be further researched, using different statistical methods (e.g., factor analysis) in a variety of settings. Third, a comparison of reliability in analytic scoring and holistic scoring or intra-rater reliability and more importantly, testing of the validity of the newly created rubric should be put under further investigation.

References

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.

Association internationale des interprètes de conférence. (1982). *Practical guide for professional interpreters*. International Association of Conference Interpreters.

Bontempo, K., & Hutchinson, B. (2011). Striving for an 'A' grade: A case study in performance management of interpreters. *International Journal of Interpreter Education*, *3*(1), 56-71.

Bühler, H. (1986). Linguistic (semantic) and extra-linguistic (pragmatic) criteria for the evaluation of conference interpretation and interpreters. *Multilingua*, *5*(4), 231-235.

Chiaro, D., & Nocella, G. (2004). Interpreters' perception of linguistic and nonlinguistic factors affecting quality: A survey through the world wide web. *Meta*, *49*(2), 278-293.

Gile, D. (1988). Le partage de l'attention et le 'modèle d'effort' en interprétation simultanée. *The Interpreter's Newsletter*, *1*, 4-22.

Gile, D. (1991). A communication-oriented analysis of quality in nonliterary translation and interpretation. In M. L. Larson (Ed.), *Translation: Theory and practice. Tension and interdependence* (pp. 188-200). John Benjamins.

Gile, D. (2001). Consecutive vs. simultaneous: Which is more accurate. *Interpretation Studies*, (1), 8-20.

Gillies, A. (2019). *Consecutive interpreting: A short course*. Routledge.

Glen, S. (n.d.). *Cronbach's alpha: Simple definition, use and interpretation*. Statisticshowto. https://www.statisticshowto.com/probability and statistics statistics-definitions/cronbachs-alpha-spss/

Glen, S. (n.d.). *Intraclass correlation*. Statisticshowto. https://www.statisticshowto.com/intraclass-correlation/

Hale, S. (2007). *Community interpreting*. Palgrave Macmillan.

Hale, S., Garcia, I., Hlavac, J., Kim, M., Lai, M., Turner, B., & Slatyer, H. (2012).

Improvements to NAATI testing report. NAATI. http://www.naati.com.au/PDF/INT/INTFinalReport.pdf

Jin, Y. (2017). Consecutive interpreting. In C. Shei & Z. M. Gao, *The Routledge handbook of Chinese translation* (pp. 321-335). Routledge.

Kalina, S. (2005). Quality assurance for interpreting processes. *Journal des traducteurs/Meta: Translators' Journal*, *50*, 768-784.

Kurz, I. (1989). Conference interpreting: User expectations. In D. Hamond (Ed.), *Coming of age: Proceedings of The 30th Annual Conference of The American Translators Association* (pp. 143-148). Learned Information Inc.

Lee, J. (2008). Rating scales for interpreting performance assessment. *The Interpreter and Translator Trainer*, *2*(2), 165-184.

Lee, S. B. (2015). Developing an analytic scale for assessing undergraduate students' consecutive interpreting performances. *Interpreting*, *17*(2), 226-254.

Lee, S. B. (2018). Scale-referenced, summative peer assessment in undergraduate interpreter training: Self-reflection from an action researcher. *Educational Action Research*, *27*(2), 152-172.

Liu, M. (2013). Design and analysis of Taiwan's interpretation certification examination. In D. Tsagari & R. van Deemter (Eds.), *Assessment issues in language translation and interpreting* (pp. 163-178). Peter Lang Edition.

Mahmoodzahed, K. (1992). Consecutive interpreting: Its principles and techniques. In C. Dollerup & A. Loddegaard (Eds.), *Teaching translation and interpreting: Training talent and experience* (pp. 231-236). John Benjamins.

Mariana, V., Cox, T., & Melby, A. K. (2015). The multidimensional quality metrics (MQM) framework: A new framework for translation quality assessment. *The Journal of Specialised Translation*, *23*, 137-161.

McNamara, C. (1999). *General guidelines for conducting interviews*. Managementhelp.org. https://managementhelp.org/businessresearch/interviews.htm

Mesa, A. M. (1997). *L'interprète culturel: Un professionel apprécié. Étude sur les services d'interprétation: Le point de vue des clients, des intervenants et des interprètes.* Régie régionale de la santé et des services sociaux de Montréal-Centre.

Moser-Mercer, B. (1996). Quality in interpreting: Some methodological issues. *The Interpreters' Newsletter*, *7*, 43-55.

Nolan, J. (2005). *Professional interpreting in the Real World series: Interpretation techniques and exercises*. Linguistic services.

O'Brien, S. (2012). Towards a dynamic quality evaluation model for translation. *Journal of Specialized Translation*, *17*, 55-77.

Ouvrard, G. (2013). L'interprétation consécutive officielle. *Traduire*, 229, 81-95.

Pienaar, M., & Cornelius, E. (2015). Contemporary perceptions of interpreting in South Africa. *Nordic Journal of African Studies*, *24*(2), 186-206.

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

Riccardi, A. (2002). Evaluation in interpretation: Macrocriteria and microcriteria. In E. Hung (Ed.), *Teaching translation and interpreting 4: Building bridges* (pp. 115-126). John Benjamins.

Roberts, R. P. (2000). Interpreter assessment tools for different settings. In R. P. Roberts, S. E. Carr, D. Abraham, & A. Dufour (Eds.), *The critical link 2: Interpreters in the community* (pp. 103-120). John Benjamins.

Russell, D., & Takeda, K. (2015). Consecutive interpreting. In R. Jourdenais & H. Mikkelson (Eds.), *The Routledge handbook of interpreting* (pp. 88-102). Routledge.

Setton, R., & Dawrant, A. (2016). *Conference interpreting: A complete course*. Benjamins.

Shuttleworth, M., & Cowie, M. (1997). *Dictionary of translation studies.* St. Jerome.

Tran, P. L., & Do, M. H. (2022, April 24). *Interpreting quality assessment criteria and implications for English-Vietnamese consecutive interpreting quality assessment in educational context* [Conference presentation abstract]. ULIS National Conference 2022, Hanoi, Vietnam.

Wang, J.-H., 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.

Wu, J., Liu, M., & Liao, C. (20l3). Analytic scoring in interpretation test: Construct validity and the halo effect. In H.-H. Liao, T.-E. Kao & Y. Lin (Eds.), *The making of a translator. Multiple perspectives* (pp. 277-292). Bookman.

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

NGHIÊN CỨU VỀ TIÊU CHÍ ĐÁNH GIÁ CHẤT LƯỢNG BÀI THI PHIÊN DỊCH ỨNG ĐOẠN ANH - VIỆT

Trần Phương Linh, Đỗ Minh Hoàng

Trường Đại học Ngoại ngữ, ĐHQGHN, Phạm Văn Đồng, Cầu Giấy, Hà Nội, Việt Nam

Tóm tắt: Sử dụng kết quả nghiên cứu của Trần và Đỗ (2022), nghiên cứu này tìm hiểu về độ tin cậy và phản hồi của người dùng với rubrics xây dựng bởi hai tác giả để đánh giá bài thi phiên dịch ứng đoạn Anh-Việt của sinh viên tại Trường Đại học Ngoại ngữ - Đại học Quốc gia Hà Nội. Năm đánh giá viên gồm 2 đánh giá viên nhiều kinh nghiệm và 3 đánh giá viên ít kinh nghiệm đã chấm mười bài thi dịch nói khác nhau một các độc lập và đưa phản hồi về rubric này. Kết quả cho thấy rubrics mới được xây dựng khá thân thiện với người dùng và có tính ứng dụng trong đánh giá dịch nói. Nhìn chung, tính thống nhất trong đánh giá giữa các đánh giá viên, thể hiện qua chỉ số Cronbach's alpha và hệ số tương quan nội bộ, cho kết quả ở mức có thể chấp nhận được. Bên cạnh đó, giá trị thu được giữa các đánh giá viên ít kinh nghiệm cao hơn đánh giá viên nhiều kinh nghiệm. Nhận thức của người đánh giá về từng tiêu chí và quy trình đánh giá có thể giải thích cho sự khác biệt trong quyết định điểm số của họ. Các phát hiện cũng đề xuất cải thiện về từ ngữ sử dụng khi mô tả từng tiêu chí, trọng số và tập huấn đánh giá viên.

Từ khóa: bài thi phiên dịch ứng đoạn, tiêu chí đánh giá, rubrics

Appendix – C1 Test Assessment Rubric

Band	Content Fidelity	Target Language Quality	Delivery
5	- Conveys a asense of original message with **complete accuracy** in which there is **no** opposite meaning and **no** unjustified change in meaning. - Transfers **all** the information **with no** omissions, additions and **no differences** in numbers and names. - **Very logically** organizes information and ideas; forms a **natural** whole.	- Demonstrates **skillful** use of vocabulary and specialist terminology. - Produces **idiomatic and on the whole correct** expressions, and **entirely appropriate** register. - Shows a **master** control of TL grammar with **very few or no** errors	- Interprets **smoothly** with **only rare** hesitation, pauses, fillers or false starts, repetition or self-correction. - Has excellent voice projection with **precise and easily understandable** pronunciation and **easily heard** volume. - Reflects intonation, emphasis and tone appropriate to situation. - Displays a **courteous and confident** manner.
4	- Conveys a sense of original message **accurately.** - Makes only a **few** minor unjustified omissions and/or additions or distortions but not affecting transfer or comprehension of information. - **Logically** organizes information and ideas; forms a **coherent** whole.a	- Demonstrates **good** use of vocabulary and specialist terminology. - Produces **general idiomatic and on the whole mostly correct** expressions, and **largely appropriate** register. - Shows a **proficient** control of TL grammar with **occasional minor** errors.	- Interprets **for most parts smoothly** with **occasional** hesitation and/or pauses, fillers or false starts, repetition or self-correction. - Has good voice projection with **mostly understandable** pronunciation and **easily heard** volume. - Reflects intonation, emphasis and tone appropriate to situation. - Displays a **few signs of nervousness**.
3	- Adequately conveys a sense of original message - Makes **some minor or one major** inaccuracy**,** unjustified omissions, additions and/or distortions. - Organizes information and ideas with **occasional awkward or oddly placed elements**.	- Demonstrates **adequate** use of vocabulary and specialist terminology. - Produces **to certain degree idiomatic and correct** expressions, and **acceptable** register. - Shows a **weak** control of TL grammar with **frequent minor** errors.	- Interprets with **some obvious but still acceptable** hesitation or pauses, **ineffective** fillers or false starts or repetition or self-correction. - **Occasionally** displays **faulty** pronunciation but **without impairing messages**. - Makes **reasonable attempts** to reflect suitable intonation, emphasis and tone. - Displays **occasional nervous habits**.

2	- **Partially** conveys a sense of original message - Makes **some serious** inaccuracies, unjustified omissions, additions and/or distortions. - Somewhat awkwardly organizes information and ideas with **frequent awkward or oddly placed elements**.	- Demonstrates **frequently inappropriate or little** use of vocabulary and specialist terminology. - Produces **unnatural** expressions, and **inappropriate** registers, **which impact on comprehension**. - Shows **some lack** of control of TL grammar with **numerous** errors.	- Interprets with **many long** pauses**, numerous ineffective** fillers or false starts and unnecessary repetition or self-correction. - Has **frequent mispronunciation** which **causes some difficulty for the listener.** - **Fails** to reflect suitable intonation, emphasis and tone. - Displays **many notable nervous habits**.
1	- **Fail to** convey a sense of original message. - Makes **many serious** inaccuracies, unjustified omissions, additions and/or distortions. - **Disorganizes** information and ideas. They do **not flow together** and are **unrelated.**	- Demonstrates **excessive inappropriate or no** uses of vocabulary and specialist terminology. There may be numerous ineffective SL inferences. - Produces expressions and registers which **may impede comprehension**. - Shows **no** control of TL grammar.	- **Lacks fluency** which may result in **low comprehensibility**. - Speech is **unintelligible** - Speaks in a **flat monotone**. - **Lacks confidence**.
0	No interpretation is produced		