KARAMANOĞLU MEHMETBEY ÜNİVERSİTESİ

ULUSLARARASI
FİLOLOJİ ve ÇEVİRİBİLİM DERGİSİ

INTERNATIONAL JOURNAL OF
PHILOLOGY and TRANSLATION STUDIES

MAKALE BİLGİLERİ
ARTICLE INFO

Geliş Tarihi / Submission Date
24.04.2021

Rapor Tarihleri / Report Dates
Hakem/Reviewer 1 - 10.05.2021
Hakem/Reviewer 2 - 11.05.2021

Kabul Tarihi / Admission Date
28.05.2021

e-ISSN
2687-5586

Künye: (Araştırma Makalesi) Sevda BALAMAN, (2021). "A Comprehensıve Revıew Of Systematıc Assessment Technıques In Interpretıng" *Karamanoğlu Mehmetbey Üniversitesi Uluslararası Filoloji ve Çeviribilim Dergisi*, C.3/1, s. 23-45.

A COMPREHENSIVE REVIEW OF SYSTEMATIC ASSESSMENT TECHNIQUES IN INTERPRETING

SÖZLÜ ÇEVİRİDE SİSTEMATİK DEĞERLENDİRME TEKNİKLERİNİN KAPSAMLI BİR İNCELEMESİ

Sevda BALAMAN
Dr. Öğr. Üyesi, Sivas Cumhuriyet Üniversitesi, Edebiyat Fakültesi, İngilizce-Fransızca Mütercim-Tercümanlık Bölümü
sevdabalaman58@gmail.com

Öz

Sözlü çeviri eğitiminde değerlendirmenin öneminin kabul edilmesi ile sözlü çeviride değerlendirme konusunun farklı bakış açıları ilegeniş kapsamlı analizlerine daha fazla ihtiyaç ortaya çıkmaktadır. Bu sebeple, bu makale, sözlü çeviri performansının eğitsel açıdan etkili bir şekilde nasıl değerlendirileceğine ışık tutmak için sağlam metodolojilere dayanan pratik değerlendirme tekniklerini derinlemesine gözden geçirmeyi amaçlamaktadır. Bu amaçla öncelikle bu araştırma, ilgili alandaki başlıca kavramlara değinereksözlü çeviride değerlendirmenin teorik temellerini sunmayı amaçlamaktadır. Bu bağlamda, bu metin sözlü çeviride değerlendirmenin ayrıntılı bir tanımlamasıyla başlamakta ve sürecin temel noktalarını, yani geçerlilik ve güvenilirliği ve amaca göre farklı değerlendirme türlerini detaylandırmaktadır. Daha sonra, bu inceleme yazısı, bütünsel değerlendirme teknikleriyle karşılaştırmalar yaparak analitik derecelendirme ölçeklerini incelemeyi hedeflemektedir. Son olarak, bu metin, ilgili alan-yazında sunulan bazı yenilikçi değerlendirme uygulamalarının, yani sözlü çeviride akran ve öz değerlendirme tekniklerinin, bu tekniklerle ilişkili farklı parametrelere değinerek,kapsamlı incelenmesine odaklanmaktadır. Bu çalışmadan çıkarılan sonuçlar, test geliştiricilerine ve sözlü çevirmen eğitmenlerine, sözlü çeviri performansını ölçmede etkili ve sağlam test tasarımları planlamada daha fazla bilgi edinmeleri yönünde fayda sağlayabilir.

Anahtar Kelimeler: Değerlendirme, Sözlü Çeviri, Akran Değerlendirmesi, Öz Değerlendirme

Abstract

With the acknowledgment of the prominence of assessment in interpreting education, there appearsa growing need for far-reaching analysesofthe assessment issue in interpretation from different perspectives. Therefore, thisarticleis intended fordeeply reviewing the practical assessment techniques grounded inrobust methodologiesin order to cast light on how to assess the interpreting performance effectively from the educational standpoints. To this end, firstly, this research aims to present the theoretical underpinnings of the assessment in interpreting by addressing the major concepts in the relevant field. In this respect, this text starts with a detailed description of assessment in interpretation and elaborates the centralpoints of the process, i.e., validity and reliability, and different assessment types by purpose. Then, this review paper has aimed atscrutinizinganalytic rating scales by making comparisons with holistic assessment techniques. Finally, this text focuses on the thorough examination of some innovative assessment practices offered in the relevant literature, i.e., peer and self-assessmenttechniquesin interpreting, by addressing different parameters in relation to these techniques.The conclusions drawn from this study might benefit test developers and interpreter trainers to gain further knowledgeaboutplanning effective and sound test designs in measuring the interpreting performance. l and to examine it in terms of mystical meanings.

Keywords: Assessment, Interpreting, Peer-Assessment, Self-Assessment

1. INTRODUCTION

In line with the profoundly growing demand for multilingual communication in the international arena, addressing the need for employing more professional interpreters (Wu, 2010), who are the mediators of languages and cultures (Sawyer, 2004), interpretation as a discipline in its own place has started to receive greater attention in the pedagogical settings (Boa, 2015). According to Niska (2005, p. 36), approximately only 230 academic institutions were offering interpreting education worldwide in over 60 countries by the end of the 20th century. But with the expanding interest in interpreting as a profession, the number of the interpreting training and education programs has substantially increased in the last decades worldwide, specifically at the graduate level (Boa, 2015). In turn, the "landscape of interpreter education" has faced major changes, as well, in company with the prevalence of such programs in the world at an accelerated rate (Boa, 2015, p. 400).

In this sense, the issue of assessment in interpreting education is no exception to the phenomenon, in that with the recognized value of the assessment in interpreting to ensure the quality of the profession, the assessment process in this field has also undergone substantial changes in the last decades (Pöchhacker, 2004). Traditionally, in its earlier stages, the assessment of the renditions is largely product-oriented, where students are solely evaluated with a focus on measuring their outputs, i.e., the quality of their interpretation performance (Iglesias Fernández, 2011), mostly ignoring the multidimensional nature of the interpreting process (Pöchhacker, 2001) in which various attributes such as linguistic, social (Wu, 2010), affective, cognitive, or psycho-motor skills operate synchronously (Doğan, Ribas, & Mora-Rubio, 2009, p. 71). However, upon acknowledging the complex nature of the interpreting system (Deysel, 2018; Hatim & Mason, 1997), shaped by not only various internal factors (Hatim & Mason, 1997), but also some external conditions that are not easy to be managed by the interpreters in most cases (Lee, 2008), such as the physical setting where the interpretation task is performed, task difficulty, or speed rate (Deysel, 2018), it has been noted that the assessment process of interpreting should be grounded in composite models which take advantage of both product and process-oriented evaluation approaches (Gile, 2001; Iglesias Fernández, 2011; Z. Lee, 2015).

The adoption of such proposed modelling necessitates the migration in assessment from the process-oriented approach in the initial stages of learning to product-oriented assessment models in the later periods(Gile, 2001; Iglesias Fernández, 2011). Accordingly, this paradigm shift might allow for more reliable and completeperformance-based evaluations (Gipps, 1994; Lee, 2011; Sawyer, 2004), covering the considerations based on not onlythe macro andmicro criteria

(Riccardi, 2002), but also the skills and the strategies adopted in the process of delivering the interpretation output (Gile, 2001; Sawyer, 2004), in addition to the scrutiny of psychological processing (Lee, 2011).

Yet because performance-based assessment types inevitably have a more subjective (Lee, 2008),impressionistic (Sawyer, 2004),and intuitive nature (Pöchhacker, 2004), in the recent years, analytic scoring frameworks have been put forward as promising tools (Lee, 2008) which can ensure the reliability and the validity parameters in testing the interpreting performance (Sawyer, 2004). With the expanding popularity of these criterion-referenced assessment types (Wu, 2010), in comparison with the holistic assessment models (Han, 2018c), whichregard the quality of interpretationholistically, not assigning varying weightings to diverse facets of quality(Lee, 2008, p. 170), a number of studies (e.g., Han, 2017; Lee, 2008; Wang, Napier, Goswell, & Carmichael, 2015) have started to test the utility of the analytic rating tools along with the pre-determined descriptors that help raters to measure interpreting performances by making objective and transparent comparisons with the levels specified in each criterion in assessing the renditions(Lee, 2008).

In addition to the administration of these scales as teacher-rating tools (Han, 2017; Lee, 2008), alternative evaluating ways have also been suggested with a purpose of enhancing the interpreting skills in the long run, by incorporating these scoring grids in peer- and self-ratings (Han, 2018a). In such practices, it has been aimed at measuring students' interpreting performance from different angles by providing them with more transparent, practical, constructive, and explicit feedback (Han, 2018c; Z. Lee, 2015). In this sense, this review paper sets out to shed new insights into the sound assessment techniques in interpretation by illustrating some alternativesfor scoring systems which test the multi-facets of the interpreting system in "objectively measurable standards" (Kalina, 2005, p. 768). To this end, this paper is intended to provide an extensive review about the assessment techniques grounded in a theoretical basis in line with the relevant literatureby offering convincing conclusions regarding how to formally assess interpreting from a pedagogical standpoint.

2. CONCEPTUALIZING "ASSESSMENT" IN INTERPRETING

Given that "conference interpreting is a testing-intensive profession" (Setton&Dawrant, 2016, p. 373), it seems unsurprising to find that the issue of how to assess the interpretation performance is integral to the interpreting profession (Sawyer, 20004) to standardize the quality of the interpretation service in the field and to efficiently administrate the professional certification

exams (Han, 2018c; Wu, 2010). Similarly, it also appears that assessment iscritical to education programs aiming at training would-be interpreters (Choi, 2006, p. 276), i.e., interpreting students, to prepare them for their future career in the interpretation profession(Sawyer, 2004; Wu, 2010).

Although the term "assessment" is often used interchangeably with other expressions such as "evaluation", "measurement" or "test" (Bachman, 1990, p. 18; Bachman, 2004), these processes are distinct, but connected concepts in practice (Bachman, 2004; Lee, 2011), which, by and large, center around the idea of making judgements in regards to the certain characteristics of a test taker (Lee, 2011). According to Lee (2011), assessment is a concept that "is an integration and a partial superset of the terms", namely "test", "measurement" and "evaluation" (p. 90).

Assessment specifically in education refers to "making a judgment about students' learning in order to identify their strengths and weaknesses, which usually involves assigning a mark or a grade to their performances and achievements" (Wu, 2010, p.3). Regarding the interpreting training programs, the assessment issue specifically aims to "diagnose students' ability, check progress, evaluate, and compare his or her interpretation performances for the purposes of course examinations and exit mechanisms" (Lee, 2011, p. 88). Historically, the previous attempts that shed light on the assessment process in interpreting education adopted a product-oriented approach (Z. Lee, 2015), in which interpreting errors were identified by assessors as part of measurement by displaying the frequency of the data (e.g., Altman, 1974; Barik, 1971), mostly in linguistic (Barik, 1971), propositional, paralinguistic or lexical basis (Han, 2017), but not much in discourse or pragmatic level of interpreting (Clifford, 2001). Such an assessment practice,also referred to the "*atomistic notion of quality"*(Iglesias Fernández, 2013, p. 51) based on identifying the errors in lists of each criterion (Han, 2017, p. 196; Iglesias Fernández, 2013, p. 51), has not been deemedreliable because of the interdependent nature of criteria laid out in the assessment tool (Iglesias Fernández, 2013).

Indeed, the product-oriented (Z. Lee, 2015) or "quality-oriented assessment" (Iglesias Fernández, 2011, p. 12) is considered more suitable for summative assessment in educational settings or certification exams administered in interpreting profession to screen if the candidates have met the required criteria specified in accordance with the professional standards. However, such an assessment type may not generate effective results in formative evaluation, especially in the initial stages of the learning (Iglesias Fernández, 2011). Therefore, there has been a call for reshaping the assessment practices (Sawyer, 2004) in a way that privileges not only the product, i.e.,the output of the interpreting performance, but also the interpreting process, i.e.,the procedure whereby the necessary skills and strategies are acquired to be able to generate an acceptable

product (Gile, 2001; Iglesias Fernández, 2011; Lee, 2011; Pöchhacker, 2001; Sawyer, 2004; Z. Lee, 2015).

In line with this call, Iglesias Fernández (2011) states that interpreting education should pursue several stages in assessment, initially evaluating the process as part of the formative phase; then, in the intermediary stage, turning towards product-oriented assessment of quality to spot the elements of the interpreting act in terms of the cognitive, pragmatic, strategic, and interactional processing; as the final stage, as expected according to the professional standards, mainly concentrating on the end product (p.12). Depending onthis, it seems that assessment techniques vary in diverse contexts for different purposes (e.g., formative or summative assessment types) (Han, 2018c).

2.1. Purposes of Assessment: Formative & Summative Assessment

Assessment is mainly categorized into two, i.e., formative and summative assessment ways, according to its purposes (Hatim & Mason, 1997). Whereas the former, *formative assessment*, adopts a process-oriented approach in which learners are informed about their strong or weak points of their competence in the learning process, the latter, *summative assessment*, is a product-oriented assessment method which provides no feedback with the learners about their learning in regards their weaknesses or strengths, (Child, 2004), but serves as a framework for making an ultimate decision about a performance (Hatim & Mason, 1997, p. 166). In other words, while the summative evaluation methods are based on 'assessment *of* learning', the formative evaluation practices rely on 'assessment *for* learning'(Gipps, 1994; Han, 2019, p. 91; Lee, 2019, p. 154). Table 1 shows the different characteristics of the two main assessment methods (Bell &Covie, 2001; Han, 2019, pp. 90-91), as follows:

Table 1: Different points between formative and summative assessment

Formative assessment	**Summative assessment**
On-going and progressive evaluation procedure	One-off and static activity
Assessment conducted during the learning process	Assessment conducted after the learning process
Stakeholders share responsibility in promoting learning	Teachers are the main responsible participants in the assessment process
Provision of feedback	No feedback

(Bell &Covie, 2001; Han, 2019, pp. 90-91)

In the case of interpreting, the formative assessment practices can be operated during the learning process through the provision of continuous feedback in order to enable student

interpreters to diagnose in which parts they have difficulty and need remedial training. Through the summative evaluation method, assessors can make judgments on students' interpreting acts at the end of the process in learning, such as in final exams, or in professional certification tests to screen interpreters' overall capabilities (Hatim & Mason, 1997, p. 166).

All in all, in either assessment practice, the critical point is that assessors need to measure the interpreting performance by ensuring two fundamental dimensions of the test-making process, namely reliability and validity constructs (Sawyer, 2004).

3. VALIDITY & RELIABILITY IN INTERPRETING TESTING

The criticality of the validity and reliability concepts has long been recognized in assessment as the central foci of the process (Gipps, 1994; Wu, 2010; Z. Lee, 2015). The former concept, *validity*, refers to "the extent to which an assessment measures what it purports to measure" (Gipps, 1994, p. vii). In other words, a test is regarded as valid when it measures what it has been originally intended to, but not the irrelevant skills or competences (Child, 2004). And if the assessment tool does not achieve this, i.e., validity, its results are misleading, or useless (Gipps, 1994, p. 58).

Out of four main sub-types of validity, namely *construct,concurrent, content, and predictive validity* dimensions (Gipps, 1994, pp. 58-59), *construct validity* is deemed as the most critical perspective for the raters in interpreting (Wu, 2010, p. 71). Because construct validity explores the underlying skills and competences which are not directly measured in the given test (Gipps, 1994), which is the case for the interpreting tasks involving multiple test constructs to be rated, it is not plausible to accept any assessment tools which fall short of construct validity as valid instruments in performance-based assessment ways, including interpreting measurements (Wu, 2010, p. 45).

As for the concept of *reliability*, it can be stated that this construct is primarily concerned with the evaluation instrument that generates "the same result for people of similar skill levels regardless of who administers the test, who rates the test, when the test is given or what version of the test is applied" (Roat, 2006, p. 9). In fact, this definition brings with itself a few important considerations in relation to test reliability (Wu, 2010). The first dimension, *intra-rater reliability*, is that an assessment tool to be viewed as reliable needs to yield the same results for the equivalent tasks on different occasions by the same examiner (Sawyer, 2004; Gipps, 1994, p. 67). For example, in an interpreting exam, whether it be a professional or an educational testing, the examiner who measures the interpreting task is expected to ensure the consistency in his/her judgments in similar cases (Wu, 2010). The second consideration, *inter-rater reliability*, is about the agreement in different assessors' ratings in regard to the same performance. When an examiner

measures an interpreting performance, s/he is expected to generate compatible marks with another examiner who assesses the same performance (Gipps, 1994, p. 69). And the final issue to consider in the test developmentis about achieving congruent results in alternative versions of a test, such as two halves of a test (Campbell & Hale, 2003; Gipps, 1994). However, splitting an assessment tool does not seem feasible for performance-based assessment tasks (Gipps, 1994), including interpreting performances (Campbell & Hale, 2003); therefore, the examiner needs to maintain consistency in the assessment instrument by choosing the source materials in the same difficulty (Sawyer, 2004), by following a systematized procedure elaborated beforehand (Wu, 2010).

As noted above, validity and reliability are the core elements of the assessment procedure, which operate concurrently, but not in isolation (Gipps, 1994; Sawyer, 2004). At this point, Wu (2010, p. 11) suggests that "a test cannot be valid if it is unreliable". Therefore, test makers need to privilege these two elements in ideal rangethroughout thetest designing procedure, by adjusting the balance in accordance with the aim of the test (Gipps, 1994; Wu, 2010, p. 56).

However, as discussed previously, it is relatively challenging to ensure validity and reliability in performance-based evaluations including interpreting assessment because of the subjective element of the judgment (Wu, 2010). Especially, having regard to the complex (Deysel, 1998; Hatim & Mason, 1997) and multi-perspective nature of the task carried in interpreting(Pöchhacker, 2001; Wu, 2010) by enacting different skills and strategies (Doğan et al., 2009; Wu, 2010), test makers are suggested being cautious in the design of the test items by determining the assessment criteria beforehand in relation to the well-established frameworks explaining what constitutes a good interpretation (Han, 2018c;Pöchhacker, 2001; Z. Lee, 2015). In specification of the relevant criteria which lay out the core elements of a high-quality interpretation output can ensure that the testing procedure becomes more valid, measurable and objective (Han, 2018c; Z. Lee, 2015).

4. WHAT MAKES A GOOD INTERPRETATION?

Defining the criteria on the quality of good interpreting has long been debated and a unified single agreement has never been achieved in this sense (Iglesias Fernández, 2011, 2013). Following the first rigorous endeavors that attempted to explore what constitutes a good interpretation through error analysis methodologies (e.g., Altman, 1974; Barik, 1971), more systematic analyses were also conducted in the field (e.g., Bühler, 1986). Notwithstanding the value of suchattempts (e.g., Bühler, 1986) for portraying the interpretation quality, there was a call

for a shift of attention in defining the criteria of good interpretation, grounded in a multi-dimensional perspective, leaving from a mere concentration on users' (listeners) or professional interpreters' expectations to including other stakeholders' points of view regarding the "ideal quality" in interpretation (Pöchhacker, 2004; Wu, 2010, p. 21).

In line with this, Pöchhacker (2001) proposes a model which sets the multi-faceted standards for interpreting by asserting that the interpreting is linguistic-oriented and performing this task is reported as a pragmatic act. In this modelling, four dimensions were highlighted as the constituents of good interpretation: accuracy as the core element surrounded by the three layers in a hierarchy of importance, namely adequate target language use, equivalency in the intended effect, and success in communicative interaction, respectively (p. 413).

The multidimensional nature of assessment in interpreting is also elaborated from differential perspectives. For example,Kalina (2005) notes that interpretation assessment should not be conducted within a single phase, but in multiple phases by proposing a wider framework aimed at assessing the quality of professional interpreting through a macro-level approach. Depending on this approach, the researcher notes the following four processes deemed essential for the interpreting assignment: *pre-process* (making preparation for the interpretation, acquiring the necessary skills, obtaining information about interpretation and contacting with others in a group), *peri-process* (preparations by collecting data on the specific assignment to be rendered just before it is performed), *in-process* (the interpreting performance and specifying the relevant conditions and factors which are important for the performance such as booth position or delivery types), and *post-process* (post-evaluation of the process) (Kalina, 2005, pp. 780-781).

Indeed, this study has important considerations on modelling the complex nature of interpreting quality theoretically from multiple perspectives (Wu, 2010). However, its main focus is on the portrayal of the multiple phases where different elements operate for good interpretation, but not on the assessment criteria of the target. Nonetheless,the suggested model hereinmight serve as a basis for evaluation of interpretation pedagogically in training programs, even though this study attempts to provide the professional standards, without any explicit reference to assessing educational attainments in the interpreting context (Z. Lee, 2015).

Indeed, the concepts of assuring professional standards and assessment in interpreting education are distinct from each other in part, but they are not thoroughly disconnected (Sawyer, 2004; Wu, 2010). Put it in another way, there exists an interdependent link between these two foci. Therefore, it is quite reasonable to align the educational assessment theory in interpreting to the

discussion of the quality of the practice in the profession to some extent (Pöchhacker, 2004; Sawyer, 2004, p. 99). In this way, it is likely for student interpreters to be informed about the expectations and the requirements of the profession and accordingly can prepare themselves for their future career in the field (Choi, 2006).

But the emphasis in the assessment of theprofessional interpreting performance is granted to the *product* quality while the educational assessment aims at prioritizing the*process* whereby the stylistically acceptable forms of outcome are generated (Hatim & Mason, 1997; Sawyer, 2004). As such, Gile(2001) underlines the importance of judging student interpreters in a different way applied to the interpreters in profession. Upon acknowledging this, it has been suggested that there needs to be enough adjustments in the application of those criteria targeting professional interpreters to student interpreters (Lee, 2008, p. 168).

Within this scope, Riccardi's (2002) work is a pioneering attempt which puts standards on assessing the interpreting quality that differentiate for professional and student interpreters. The researcher proposed the macro-criteria to be used for professional interpretation assessment around four themes, i.e., equivalence, accuracy, appropriateness, and usability (p. 118) and micro-criteriatargeting for the student-interpreters, such as phonological morphological lexical or semantic deviations,to be applicable in both consecutive interpretingand simultaneous interpreting modes. These micro-criteria are defined in light of the interpreter trainers' evaluations and information gathered by consulting the relevant literature and the researcher's personal experience (Riccardi, 2002).

Even though Riccardi (2002) hasprovidedimportant modellings for didactic purposes in defining good quality interpretation, it is not without criticism. In this sense, Lee (2008) posits that this study does not have enoughexplanations regarding the potential methods with which the interpretation quality is translated into a numerical count(p. 168), posing an urgent need for assessing interpreting performance through systematic rating scales (Han, 2017).

Although there is a lack of single universal testing grid which is unanimously accepted by all test developers or educators in the interpreting field (Deysel, 2018; Iglesias Fernández, 2013; Wu, 2010), whether it be for professional or educational interpreting assessment, there has been some consensus on the idea that a good interpretation output covers a few broad criteria, whose "terminology may vary from one author or text to the other" (Pöchhacker, 2001, p. 413). The most widely-acknowledged criteria for assessing the quality of interpretation are summarized, as follows:

Table 2: The most-frequently used criteria for defining good interpretation in the relevant literature

	The most widely-acknowledged criteria	**Relevant previousstudies that include these criteria**
Criterion on the *correctness* of the interpretation in five different wordings that center on the same point	accuracy	Choi, 2006; Lee, 2008; Pöchhacker, 2001; Wu, 2010; Z. Lee, 2015
	loyalty	Pöchhacker, 2001
	content	S.-B. Lee, 2015; Han, 2015; Riccardi, 2002
	information completeness	Han, 2015
	sense consistency	Bühler, 1986
Criterion on the evaluation of the *language quality* used in interpretation	target language quality	Han, 2015; Lee, 2008; Riccardi, 2002; Wang et al., 2015; Z. Lee, 2015
Criterion on the *speaking/ presentation skills*	delivery features	Lee, 2008; Riccardi, 2002; Wu, 2010; Wang et al., 2015
	delivery fluency	Han, 2015

In sum, considering the points elaborated in Table 2, a suggested framework for assessing the interpretation quality for pedagogical purposes might be based on these frequently-mentioned components along withthe detailed descriptors around these indicators. In this regard, these criteria can be thought as the majorunitsof high-quality interpretationwith multiple perspectives, from the aspects of accuracy of the output, the quality of the language used in rendering the interpretation utterance, and the interpreter's speaking and presentation skills. In line with these dimensions, different scoring scales seeking to ensure validity and reliability in optimum ranges have been formulated to assess and gauge the interpreting quality (Wang et al., 2015) in a more measurable way (Choi, 2006, p. 278).

5. RATING SCALES IN INTERPRETING

To begin with, the use of *holistic scales* is one way of judging interpretation quality. This type of scoring schemes is applied when examiners are expected to rate the interpreting act "as a *whole* entity" (Han, 2018c, p. 67). The overall impression about the whole interpretation quality

is obtained by evaluating certain aspects of the performance through descriptor-oriented rating guides. Scalar descriptors necessarily reflect the typical profiles of the performance, without any explicit or transparent reference to the better performance in relevant specifications of the broad criteria. Moreover, holistic scales assume that the examinees perform all heterogenous descriptors simultaneously. Yet, the descriptors in holistic scales do not necessarily differentiate the examinees who may reflect certain characteristics of the task variably and unevenly because of the lack of the numerical scoring distributed to each criterion (Han, 2018c).

Therefore, *descriptor-based, criterion-referenced analytic scales* have started to receive greater attention from test developers and interpreting trainers to evaluate varied profiles of the test takers based on a range of codified assessment criteria along with their well-defined descriptors (Han, 2018c; Lee, 2008; Wang et al., 2015). Various dimensions of the interpretation task are graded separately by using the specifications of the given indicators with their descriptors on a band scheme (Lee, 2008; Wang et al., 2015). In these scoring grids, performance descriptors are firmly established on score bands. Examiners evaluate each domain of the interpreting task, by pursuing a prescribed cluster of descriptors, and end up with a "*score profile*" (Han, 2018c, p. 67). The composite score representing the quality of interpretation can be obtained through the weighted sum of the diverse parts of the performance by valuing each criterion equally or differentially (Han, 2018c, p. 68).

For example, while Cheung (2007) rated each criterion with equal weighting, Lee (2008) and Wang et al. (2015) measured the interpretation quality through variably distributed weightings in regard to the given criteria. Lee (2008) analyzed whether analytic rating scales could function properly in assessing interpreting performance in consecutive interpreting mode. The scale falls into three main criteria, to each of which differential weightings were distributed, as follows: 40% of weightings for accuracy and 40% for target language quality, and 20% for delivery (p. 171). The data were collected through eliciting both professional and student interpreters' ratingsof the renditions in consecutive interpreting by using this three-dimension scale. Overall, the feedback yielded from the participants in the study suggests that such a scoring scale can be a promising instrument to measure the performance in consecutive interpretation (Lee, 2008).

Wang et al. (2015) also investigated the incorporation of analytic rating scales to mark the performance in simultaneous interpreting in the bidirectional functionality of American Sign Language and English. The researchers obtained the data from the two rubrics based on the four assessment criteria as the indicators of interpretation quality. Partly informed by the schemata of the weightings assigned to the criteria and the suggestions of the participants in Lee's (2008)

study,Wang et al. (2015) distributeddiverse weighting to each criterion: 50% for accuracy, 20% for target text features, 15% for delivery features, and 15% for the processing skills (p. 86).

In short, irrespective of the discussion on assigning differential or equal weightings to the given criteria in analytic tools, it has been concluded that descriptor-cased, criterion-referenced analytic tools appear to be "reliable, valid, and practical" tools (Han, 2017, p. 197) for performance assessment in interpretation. In this sense, having recognized the potential of these instruments, there is a growing interest intaking advantage of such promising scales in alternative assessment modes to enlarge the horizons towards the innovative interpreting evaluation (Han, 2018a).

6. ALTERNATIVE ASSESSMENT PRACTICES: PEER AND SELF-ASSESSMENT METHODS

With the shift in interpreter training from teacher-centered practices to student-led methodologies recently (Z. Lee, 2015), analytic rating tools receive expanding emphasis from test developers and interpreter educators to enable students to make judgements regarding their own or peers' interpreting tasks. To start with, peer assessment (PA) applications have been gaining momentum in interpreter education in the recent years (Han, 2018b) and have been accepted as one of the most widely-used methods in formative assessment practices (Han, 2018a). PA refers to "an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status" (Topping, 1998, p. 250). It has been well-documented that PA benefits students in a number of ways, such as enabling students to identify their strong or weak points and pinpoint the areas that need greater emphasis, facilitating metacognition (Topping, 2009, p. 20), gaining an in-depth awareness of assessment domains, assuming more responsibility for learning, and promoting reflectivity (Fowler, 2007; Topping, 1998).

Recognizingits value in education, PA has started to be administered in interpreting training programs of higher education with a growing interest. For example, Lee (2017) sought to reveal the perceptions of three students towards PA practices conducted within summative assessment in consecutive interpreting by using analytic rating frameworks. The data revealed that participating students had initial concerns about their unfamiliarity with the implementation, in that they were not quite sure about whether they had the competence for applying PA accurately and thoroughly. But in time, as they got familiar with the process, the participants stated that they felt more empowered in implementation and demonstrated more responsibility.Overall, students reported that PA was a useful and reliable method for interpreting assessment. Lee (2019) conducted

another study in an action research design with an aim for investigating the researcher-teacher's self-reflections on the incorporation of scale-based summative peer ratings utilized in consecutive interpreting. The findings obtained from different instruments demonstrated that PA had the credibility and benefits, despite its procedural difficulties specific to the local context (Lee, 2019, p. 152).

Unlike Lee's (2017, 2019) research targeting summative assessment, Han (2018b) conducted an empirical study analyzing whether formative peer evaluations performed in consecutive interpreting were accurate and trustworthy. It was illustrated in the study that raters' scorings based on an analytic scale with three dimensions of quality varied in the accuracy domain. In addition to this, it was revealed that peer assessors yield more accurate ratings in certain circumstances, that is, while evaluating certain students in comparison to others, assessing target language quality compared to other quality dimensions and rating the performance in the directionality of English-to-Chinese interpretation rather than the opposite way (Han, 2018b, p. 979).

Similarly, Han and Zhao (2020) carried out an evidence-based study intended to test the accuracy of students' gradings of their peers with the help of a rating scale in a two-way directionality of interpretation. The researchers documented that English-to-Chinese interpretation was rated more easily than the other direction. Moreover, fidelity was assessed more accurately as compared to fluency and expression. And experienced assessors were more accurate in grading than novice counterparts. Su (2019) also examined students' peer ratings as to the interpretation quality in a simultaneous interpreting class in English-to-Chinese direction. Having trained about how to implement PA by operating a scale established on three components of interpretation quality, i.e., "accuracy, presentation and target language quality" (p. 177), participants graded three samples of interpretation tasks along with their reflection on three criteria of the construct, namely interpretation quality. The findings indicated that there were more evaluations about deviations in the target language structure and fluency in their presentations as compared to the accuracy domain (Su, 2019).

As seen from the above, it is apparent that a range of factors seem to have an imposing effect on the efficiency of PA, such as the domain of the assessment, the direction of the interpretation or the raters' characteristics (Han & Zhao, 2020), suggesting that when designing such alternative methods in interpreting classrooms, test developers and educators be cautious about these points (Han, 2018a).

As for the other innovative assessment mode, self-assessment (SA), it can be noted that given "interpreter training courses are intensive in nature and training is complemented by additional self-study hours" (Deysel, 2018, p. 27), it is unsurprising to document that the notion of self-evaluation has taken relative attention in interpreting training (Han &Fan, 2020), as a critical component of both formative (Han & Fan, 2020; Han &Riazi, 2018) and ipsative assessment which refers to the process where students carry out evaluations regarding their performance through making comparisons with their previous performances (Gipps, 1994, p. vii). The implementation of SA, deemed integral to the interpreting process, is based on the idea that students carry out their self-evaluations about different aspects of their learning (PostigoPinazo, 2008).

It has been stated that SA brings substantial benefits to the interpreting trainees such as promoting self-criticism (PostigoPinazo, 2008), fostering self-reflectivity, self-regulation (Han & Fan, 2020; Han &Riazi, 2018), metacognition (Lee, 2011), and autonomy (Han & Fan, 2020), gaining in-depth awareness of the interpreting task (Fowler, 2007), recognizing strong or weak sides (Han & Fan, 2020; Han &Riazi, 2018), cultivating motivation, active participation, and individual and group self-esteem, increasing students' responsibility for the learning (PostigoPinazo, 2008, p. 197), and enhancing professional growth (PostigoPinazo, 2008) mediated by sustainable assessment critical to life-long learning (Han & Fan, 2020; Han &Riazi, 2018).

Within this context, a small but expanding body of research has attempted to examine the utility of SA in interpretation (e.g., Choi, 2006; Fowler, 2007; Han & Fan, 2020; Han &Riazi, 2018; PostigoPinazo, 2008) and to what extent trainees can self-assess their interpreting performances accurately by utilizing analytic rubrics has become the central research topic for researchers (e.g., Han &Riazi, 2018; PostigoPinazo, 2008). This research theme echoes in a longitudinal study by Han and Riazi (2018), who sought to analyze the extent to which students accurately rated their interpretation performances and in what ways the accuracy level of SA would change in the course of time. Data showed that "general accuracy" of SA (p. 386) improved in time for both English-to-Chinese and for the other way, even though their study yielded a weak-to-moderate correlation between student's self-ratings and teacher ratings in diverse components of the interpretation such as target language quality or fluency (Han &Riazi, 2018, p. 394).

Out of the earlier studies, few of them also foregrounded the interpreting students' reflections about the SA process, having participated into SA-oriented practices (e.g., Deysel, 2018; Han & Fan, 2020). In this sense, Han and Fan (2020) explored the effectiveness of the SA process through students' lens by reporting 38 interpreting students' reflections on the SA-oriented implementations carried out in a 10-week process as part of consecutive interpreting course. The

qualitative data displayed that the overall perception of the SA-facilitated process was useful and this kind of assessment appeared as promoting the concepts of self-awareness, self-monitoring and self-reflectivity.In sum, considering the positive sides of these alternative assessment practices, it can be noted that thesemethodologies can inspire educators to adopt an interpreting assessment process where learners are not only subjected to the testing procedure, but also equipped with strategies for improving themselves in the learning-to-interpret process.

7. CONCLUSION

This review study is aimed to examine the assessment issue in interpreting which occupies a central position in this field of study for both professional and trainee interpreters (Wu, 2010). To this end, this paper, firstly, has presented the theoretical foundation of the major concepts in assessment by indicating different functionalities of the variable. In this sense, this review text has unpacked the assessment issue by discussing what constitutes a good interpretation. The most-widely adopted criteria utilized fordefining the quality of interpretation(see Table 2) can be the starting point for educators and test designers indeveloping test items. Hence, it could be possible to measure the interpretation output from different dimensions globally agreed on.

Importantly, these core constituents of the good-quality interpretation should be covered in the well-defined analytic scoring frameworks which seem to be more functionalin comparisonwith the holistic assessment ways, as discussed in Section 5. The analyticscalesshouldbe designed in such a way that reasonable weightings are distributed to the relevant criteria. At this point, test developers, first and foremost, need to specify the purpose of the assessment and then make logical considerations in weighting which will be assigned to each criterion, in agreement with the aim of the evaluation. Hence, the assessment process can yield true and well-groundedgradings, which seems a highly challenging issue to achieve in performance assessment practices, as in the case of interpreting activities (Wu, 2010).

In addition to this, the current text has also concentrated on the alternative assessment methodologies, i.e., peer- and self-ratings, within the scope of both formative and summative assessment practices (Han, 2018a).However, these assessment ways are not without limitations. First among them is the inherent subjectivity of the PA and SA types. That is, students might tend to overmark the self or peers' performances, hindering the accuracy of the ratings (Han, 2018a, 2018b). In this sense, the reliability and validity facets appear as the critical parameters which might impact the value of these assessment modes (Han, 2018a). However, under the thorough guidance, students can be scaffolded about how to properly perform peer (Lee, 2017, 2019; Wang

et al., 2015) and self-rating implementationsthrough teachers' constructive exemplars (Han, 2018a; Han & Fan, 2020). At this point, encouraging students to take advantage of scoring grids builton objective, transparent, and explicit criteria might seem prominent (Han & Fan, 2020; Z. Lee, 2015).

Moreover, producing numeric scores should not constitute the primary focus of alternative assessment practices (Han &Riazi, 2018). Instead, the undeniable importance of such evaluation ways resides in their potentials for promoting self-awareness of the process and monitoring their own performance by indicating their weak or strong points, acknowledged as vital elements ofsustainable growth in interpretingfor professional life(Fowler, 2007).

Second is thetime-consuming nature of both SA and PA implementations. Since the application of such models inevitably consumes a large portion of class time (Han & Fan, 2020), trainers can have a tendency to ignore the criticality of these evaluation ways. But when students are exposed to the ratings of the relevant performances through SA and/or PA ways on repeated occasions, it is possible for them to have broadened insights into the proper implementations of the processes (Fowler, 2007), which might in turn save time for trainers in applying both formative and summative evaluation practices (Han, 2018a).

In addition, it would be exhausting to conduct such methods in each assessment session from both teachers' and learners' perspectives. Therefore, a reasonablebalance should be maintained between teacher ratings and alternative assessment practices, i.e., peerand self-evaluation modes,ingrading the interpreting outputs.Without ignoring the value of teacher feedback and measurement, educators can complement teacher ratings with these innovativetechniques when possible, as extensions of the learning/teaching process. It should not be overlooked that these assessment techniques may not always produce fruitful outcomes,considering that the efficiency of theseways largely depends on their systematic implementations in the learning process. In line with this, after each application of such practices, educators should take student interpreters' reflections on the implementation of these methodologies in order to specify the weak sides of the procedures with the intention of eliminating them for the next sessions.Overall, if these techniques are integrated into the assessment process within a careful and organized design, in agreement with the curricular objectives, it is apparent that these two models can ensure the innovation of traditional assessment procedure as promising tools, in consistent with the modern-day interpretation training programs which indicate a change towards student-led orientations (Z. Lee, 2015).These evaluation modes, strengthened by the incorporation of analytic rating tools in well-defined criteria beforehand, instead of holistic rating schemes, will probably allow student

interpreters to both improve themselves and be tested synchronously. Being evaluated through the lens of their own or peers, will likely enable these students to obtain constructive feedback and in turn might help them to see the testing session as an opportunity for self-improvement, a vital part of the life-long learning process (Fowler, 2007). In conclusion, this text, specifically based on reviewing the viable assessment methodologies in interpretation, is intended to benefit educators and test developers by helping them to broaden the perspectivesfor carrying out the performance-based evaluations of interpreting tasks, in particular.

REFERENCES

Altman, J. (1994) Error analysis in the teaching of simultaneous interpretation: A pilot study. In S. Lambert &B. Moser-Mercer (Eds), *Bridging the Gap. Empirical research in simultaneous interpretation (*pp. 25–38). Amsterdam/Philadelphia: John Benjamins.

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

Bahman, L. F. (2004). *Statistical analysis for language assessment*. Cambridge: Cambridge University Press.

Bao, C. (2015). Pedagogy. In R. Jourdenais& H. Mikkelson (Eds.), *The Routledge handbook of interpreting* (pp. 400-416). New York, NY: Routledge.

Barik, H. C. (1971). A description of various types of omissions, additions and errors of translation encountered in simultaneous interpretation. *Meta, 16*(4), 199-210.

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, *85*(5), 536–553.

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

Campbell, S. & Hale, S. (2003). Chapter 15. Translation and Interpreting Assessment in the Context of Educational Measurement. In G. Anderman & M. Rogers (Eds.), *Translation Today* (pp. 205-224). Bristol, Blue Ridge Summit: Multilingual Matters. https://doi.org/10.21832/9781853596179-017

Cheung, A. K. -F. (2007). The effectiveness of summary training in consecutive interpreting (CI) delivery. *Forum,* 5(2), 1–23.

Child, D. (2004). *Psychology and the Teacher (7th ed.)*. London, New York: Continuum.

Choi, J. Y. (2006). Metacognitive evaluation method in consecutive interpretation for novice learners. *Meta,51*(2), 273–283. https://doi.org/10.7202/013256ar

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

Deysel, E. (2018). *Self-assessment by computer-assisted interpreter training (CAIT) for practicing interpreters: parliament as a case study* (Unpublished Master's Thesis). Stellenbosch University, Stellenbosch, South Africa.

Doğan, A., Ribas, M. A., Mora-Rubio, B. (2009). Metacognitive tools in interpreting training: A pilot study. *HacettepeÜniversitesiEdebiyatFakültesiDergisi (Hacettepe University Journal of Faculty of Letters), 26*(1), 69-84.

Fowler, Y. (2007). Formative Assessment: Using Peer and Self-Assessment in Interpreter Training. In C. Wadensjö, B. E. Dimitrova, & A-L. Nilsson (Eds.), *The Critical Link 4: Professionalisation of Interpreting in the Community*, (pp. 253–262). Amsterdam: John Benjamins.

Gile, D. (2001). L'évaluation de la qualité de l'interprétationencours deformation. *Meta, 46*(2), 379–393. https://doi.org/10.7202/002890ar

Gipps, C. (1994). *Beyond testing: Towards a theory of educational assessment*. London: Falmer Press.

Han, C. (2015). Investigating rater severity/leniency in interpreter performance testing: A multifaceted Rasch measurement approach. *Interpreting, 17*(2), 255-283.

Han, C. (2017). Using Analytic Rating Scales to Assess English/Chinese Bidirectional Interpretation: A Longitudinal Rasch Analysis of Scale Utility and Rater Behavior. *LinguisticaAntverpiensia New Series –Themes in Translation Studies 16*, 196–215.

Han, C. (2018a). A longitudinal quantitative investigation into the concurrent validity of self and peer assessment applied to English-Chinese bi-directional interpretation in an undergraduate interpretingcourse. *Studies in Educational Evaluation, 58*,187–196. https://doi.org/10.1016/j.stueduc.2018.01.001

Han, C. (2018b). Latent trait modelling of rater accuracy in formative peer assessment of English-Chinese consecutive interpreting. *Assessment & Evaluation in Higher Education 43*(6): 979– 994. https://doi.org/10.1080/02602938.2018.1424799

Han, C. (2018c). Using rating scales to assess interpretation: Practices, problems and prospects. *Interpreting, 20*(1), 59 – 95. https://doi.org/10.1075/intp.00003.han

Han, C. (2019). Conceptualizing and operationalizing a formative assessment model for English-Chinese consecutive interpreting: a case study in an undergraduate interpreting course. In Huertas-Barros, E., Vandepitte, S., & Iglesias-Fernández, E. (Eds.), *Quality Assurance and Assessment Practices in Translation and Interpreting* (pp. 89-111). IGI Global. DOI: 10.4018/978-1-5225-5225-3.ch004

Han, C. & Fan, Q. (2020). Using self-assessment as a formative assessment tool in an English-Chinese interpreting course: student views and perceptions of its utility. *Perspectives, 28*(1), 109-125. https://doi.org/10.1080/0907676X.2019.1615516

Han, C. &Riazi, M. (2018) The accuracy of student self-assessments of English-Chinese bidirectional interpretation: a longitudinal quantitative study. *Assessment & Evaluation in Higher Education, 43*(3), 386-398, https://doi.org/10.1080/02602938.2017.1353062

Han, C. & Zhao, X. (2020): Accuracy of peer ratings on the quality of spoken-languageinterpreting.Assessment&Evaluation in Higher Education, https://doi.org/10.1080/02602938.2020.1855624

Hatim, B., & Mason, I. (1997). *The Translator as Communicator*. London & New York: Routledge.

Iglesias Fernández, E. (2011). Under examination. do all interpreting examiners use the same criteria? *The Linguist, 50*(2), 12-13.

IglesiasFernández, E. (2013). Unpacking delivery criteria in interpreting quality assessment. In D .Tsagari&R. vanDeemter (Eds.), *Assessment issues in language, translation and interpreting* (pp.51- 56). Frankfurt: Peter Lang.

Kalina, S. (2005). Quality assurance for interpreting processes. *Meta, 50* (2), 768–784.

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 interpretingperformances.*Interpreting,*17(2), 226–254.https://doi.org/10.1075/intp.17.2.04lee

Lee, S.-B. (2017). University students' experience of 'scale-referenced' peer assessment for a consecutive interpreting examination. *Assessment & Evaluation in Higher Education, 42* (7), 1015–1029. doi:10.1080/02602938.2016.1223269

Lee, S.-B. (2019). Scale-referenced, summative peer assessment in undergraduate interpreter training: self-reflection from an action researcher. *Educational Action Research, 27*(2), 152-172.

Lee, Y.-H. (2011). Comparing self-assessment and teacher's assessment in interpreter training. *T&I Review, 1*, 87–111.

Lee, Z. (2015). *The reflection and self-assessment of student interpreters through logbooks: a case study* (Unpublished Doctoral Dissertation). Heriot-Watt University, Edinburgh.

Niska, H. (2005). Training interpreters: Programmes, curricula, practice. In M. Tennent (ed.), *Training for the New Millennium: Pedagogies for Translation and Interpreting*(pp.36–64*)*. Amsterdam/Philadelphia: John Benjamins.

PostigoPinazo, E. (2008). Self-assessment in teaching interpreting. *TTR (Traduction, terminologie, rédaction),21*(1), 173–209. https://doi.org/10.7202/029690ar

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*. Shanghai: Shanghai Foreign Language Education Press.

Riccardi, A. (2002). Evaluation in interpreting: Macrocriteria and microcriteria. In E. Hung (Ed.), *Teaching Translation and Interpreting 4* (pp. 115-126). Amsterdam & Philadelphia: John Benjamins.

Roat, C. E. (2006). *Certification of health care interpreters in the United States: A primer, a status reportandconsiderations for national certification*. Los Angeles, CA: The California Endowment.

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

Setton, R. &Dawrant, A. (2016). *Conference interpreting: a trainer's guide*. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/btl.120

Su, W. (2019). Interpreting quality as evaluated by peer students. *The Interpreter and Translator Trainer, 13*(2), 177–189. doi:10.1080/1750399X.2018.1564192

Topping, K. 1998. Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276. doi:10.3102/00346543068003249

Topping, K.(2009). Peer Assessment.*Theory into Practice,48*(1), 20–27. doi:10.1080/00405840802577569.

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. https://doi.org/10.1080/1750399X.2015.1009261

Wu, S. C. (2010). *Assessing simultaneous interpreting: A study on test reliability and examiners' assessment behavior* (Unpublished Doctoral Dissertation). Newcastle University, the UK.

STRUCTURED ABSTRACT

The prevalence and the critical role of the interpreting profession in the world in order to ensure the multilingual interaction between different stakeholders have led to the growing number of interpreting education and training institutions worldwide (Boa, 2015). As such, the issue of the assessment in interpretation has taken more attention in order to set standards onthe quality of interpretation in both interpreting profession and training settings (Sawyer, 2004).

The term "assessment" in interpreting education provides the basis for gathering information regarding the functionalities such as identifying learners' competencies, reviewing the possible improvement, examine, and make a comparison between the learner's own performances during and at the end of the process (Lee, 2011, p. 88). Previously, the assessment in interpretation is mainly product-oriented, in which students are generally evaluated with a focus on measuring their end-products, namely their interpretation performance (Z. Lee, 2015). And such evaluations largely underestimate the multidimensional nature of the interpreting process (Pöchhacker, 2001; Wu, 2010), where an array of other factors such as linguistic, social (Wu, 2010), affective, cognitive, or psychomotor skills runsimultaneously (Doğan et al., 2009, p. 71). However, the recent efforts have been mostly attempted to take advantage of both product- and process-oriented evaluation approaches in the last decades (Gile, 2001; Iglesias Fernández, 2011; Z. Lee, 2015). These composite models, structured by the combination of these two approaches (Z. Lee, 2015), serve as a basis for more comprehensiveperformance-based evaluations (Gipps, 1994), addressing not only the macro and micro elements (Riccardi, 2002), but also the skills and strategies utilized in the interpreting process (Sawyer, 2004).

But because performance-based assessment types often yield subjective evaluations (Lee, 2008), the utility of analytic scoring frameworks has attracted extensive interest in the recent years (Lee, 2008). These scales have been formulated by different researchers (e.g., Han, 2015; Lee, 2008) by encompassing diverse criteria of good interpretation. In line with this, defining the assessment criteria regarding the quality of interpretation has become central for the research targetingnot only professionals, but also student interpreters (e.g., Riccardi, 2002).

From didactic points of view, sound assessment practices taking advantage of analytic scoring schemes have been offered by the previous empirical research conducted in different

contexts. For example, a small but growing scope of research has been carried out with a purpose ofimplementing alternative waysfor assessing the interpretation performance such as peer (e.g., Han, 2018b; Lee, 2017; Su, 2019) or self-evaluation methodologies (e.g., Han & Fan, 2020; Han &Riazi, 2018; PostigoPinazo, 2008), as part of both formative and summative assessment practices (e.g., Han, 2018a).In light of the relevant literature, this paper has aimed to review the formal assessment techniques in interpretation for educational purposes. To this end, first of all, this review paper is intended to demonstrate the theoretical specifics of the assessment in interpreting by describing the central conceptsin this field of study. Then, this paper concentrates on displaying some promising assessment practices offered by the previous research conducted in different contexts. The resulting information of this text might help those who would like to gain deeper understanding into developing effective and soundtechniquesfor measuring the interpreting performance.