http://dx.doi.org/10.30827/sendebar.v30i0.8512
ISSN-e 2340-2415 | Nº 30 | 2019 | pp. 221-244

A Rasch-based Validation of the Evaluation Rubric for Consecutive Interpreting Performance

Foroogh Khorami | Ghasem Modarresi*
foroughkhn@gmail.com | qasem.modarresi@gmail.com
Department of English, Quchan Branch, Islamic Azad University, Quchan, Iran
*Corresponding author: qasem.modarresi@gmail.com

Recibido: 09/01/2019 | Revisado: 20/04/2019 | Aceptado: 29/07/2019

Abstract

The present study is aimed at developing and validating a rubric for consecutive interpreting performance. Firstly, the researchers interviewed with interpreting experts and teachers to identify the factors involved in assessing student performance on consecutive interpreting. Then they designed an interpreting evaluation checklist including 38 items based on the commonalities that emerged from the interviews. Rasch Measurement determined that 25 items of the checklist fitted the Rasch model. Following this, the researchers constructed 25 items in a Likert-type scale with four levels. Having employed the rubric to measure 105 homogeneous interpreting student performance on consecutive interpreting, Factor Analysis revealed the presence of seven factors with eigenvalues greater than 1.0, which accounted for 77% of the variance. At the final stage, SEM results indicated a good fit to the data. The final rubric consisted of four factors with 22 items. The study offers practical implications for interpreting students and teachers.

Key words: consecutive interpreting, evaluation rubric, interpreting performance, validation

Resumen

Validación en base a Modelo Rasch, como criterio para la valoración del rendimiento de la interpretación consecutiva

La presente investigación tiene por objeto elaborar y validar un criterio de evaluación del rendimiento de la interpretación consecutiva. Al principio, los investigadores mantuvieron una entrevista con los expertos y profesores de interpretación a fin de conocer los ítems involucrados en la evaluación de la traducción oral consecutiva, y luego, crearon una lista de verificación bajo el título de valoración de interpretación, comprendida por 38 ítems de temas comunes recogidos en las entrevistas realizadas, y se descubrió que unos 25 ítems fueron compatibles con el modelo Rasch. Entonces, los investigadores redactaron 25 ítems en 4 niveles tomando como base la escala Likert, y aplicaron este criterio para la evaluación de la interpretación consecutiva de 105 estudiantes homogeneizados desde el punto de vista de conocimiento lingüístico, y como consecuencia, se reveló que este criterio cuenta con 7 factores con un valor superior a 1 y su contenido abarca un 77% de la variación. En la fase final del ensayo, los resultados del Modelo de Ecuaciones Estructurales (SEM, por sus siglas en inglés) demostraron la buena compatibilidad de los datos. El criterio de evaluación final comprendía 4 factores y 22 ítems. El presente estudio también presenta estrategias prácticas para los estudiantes y profesores de traducción oral.

Palabras clave: interpretación consecutiva, matriz de evaluación, rendimiento de interpretación, validación

1. Introduction

Regarding the contribution of language testing and assessment in any educational setting, it would be enough to assert that teachers generally spend a minimum of one-third of their instructional time on assessment-related activities (Bachman, 2014). The point is that providing effective instruction and promoting student learning appear to be closely related to the quality of assessment techniques used in the classrooms. The research works on the interplay between different disciplines would support the new findings in the field of interpreting studies. Attention to assessment criteria employed by the teachers should be at the foreground while working on consecutive interpreting with students in the classroom settings. According to Kutz (1994), interpreting competence entails different skills and knowledge, organized hierarchically from general to specific, on the basis of which certain prototypical patterns of behavior evolve with the view of solving problems.

Wilss (1998) remarked that, in research methodology, like most bipolar issues, qualitative /quantitative distinction is a matter of subjectivity vs. objectivity. Although previous theoretical studies on translation assessment focused on objective evaluation (e.g. Newmark 1988; Wilss 1998; House 1997), it is hard to carry out accurate objective evaluation. Indeed, the absence of direct observation and description of personal, social, and discoursal factors of translation is considered as the major reason (Beeby 2000). The satisfactory solution to this problem is to employ an inclusive and systematic approach to shelter the representations of these factors (as much as possible) by means of standardized tools, and to handle and take into account all of these factors appropriately in order to evaluate them in a valid and reliable way (Beeby 2000).

Stobart and Gipps (1997) view assessment as an integral part of the training process in interpreter education. Considering interpreting assessment, rubrics allow for more systematic and holistic grading (Angelelli 2009). A rubric generally contains all sub-components that constitute the underlying constructs. According to Angelelli (2009: 39), rubrics provide «descriptive statements of behaviors that candidates may exhibit in a particular sub-component». Since a scoring rubric can be used to holistically score any product or performance (Moss & Holden 1988; Walvood & Anderson 1998), it makes sense to discuss its feasibility for scoring interpretation.

Indeed, as commented by Walvood and Anderson (1998), a rubric is developed by identifying what is being assessed (i.e. translation competence). Generally, evaluation rubrics are employed in language testing and assessment to measure primary and multiple traits, or competencies, in language production (Cohen 1994). The concept of *trait* refers to a particular ability or competency that is being measured (Bachman 2014). The problem, hence, is that there have been few developed rubrics in the field of interpreting studies so far. Angelelli (2009) developed a rubric for interpreting assessment with a focus on contextualization cues and turn taking competence. This scale goes from 1 to 4, representing four levels of achievement as superior (highest

level of performance), advanced, fair, and poor (lowest level of performance). However, the validity of the rubric is under question since it is limited to performance at two levels of contextualization cue and turn taking.

In the current practice of interpreting assessment, the challenges of validity and reliability issues are in serious need of clarifications (Campbell & Hale 2003; Hatim & Mason 1997; Sawyer 2004). The problem is that in developing an analytic rubric for performance-based interpreter assessment, the essential competencies for effective interpreter performance are generally underrated. Actually, they should be identified and defined, or operationalized (Bachman & Palmer 1996), and a rating scale is needed to be used for grading each of them separately (Mertler 2001). Research into the development of scoring rubrics used for the measurement of interpreting competence is still at the initial stage of exploration in interpreting studies. As rightly declared by Muñoz (2012: 170), «we gained much insight into translators' mental life, but there has been very little construct-validating research». Indeed, understanding how examiners make judgment and developing effective test instruments based on valid test constructs are essential grounds for both practical examination administration and research studies on the issues surrounding the interpreting assessment. Since interpreting competence is interwoven with the dynamic nature of target discourse community, it is not an easy task to figure out the variables that determine if an interpreting performance is acceptable or not. As Angelelli and Jacobson (2009: 3) noted, «few researchers have focused on the measurement of aspects of interpreting and on the problems of assessing interpreting via the implementation of valid and reliable measures based on empirical research».

The important point is that by developing a scoring rubric, graders can score all the elements that are relevant to a test. This process confirms the evidence that the constructs which are intended to be measured are not only measured by the test (as a result of careful development) but also scored by graders (Wiggins 1998). The researchers of the current study decided to develop a rubric for assessing consecutive interpreting performance based on the linguistic, cultural, situational and psychological factors existed in the Iranian context. The present study opted for consecutive interpreting through which «the interpreter listens to a speech segment for a few minutes or so, takes notes, and then delivers the segment sentence by sentence or paragraph by paragraph in the target language» (Gile 1995: 42). Consecutive interpreting is easier for interpreting students than simultaneous interpreting which, as a complex task, requires listening and speaking concurrently (Mizuno 2005).

Thus, the study aimed to gather data for the development of an evaluation rubric for consecutive interpreting assessment based on the viewpoints of interpreting experts and teachers. Moreover, the study examined the actual use of the rubric in interpreting courses at the university level. The rubric can also be utilized as an objective scoring instrument to measure consecutive interpreting performance in the research domain.

2. Literature review

2.1. Testing and assessment

Knowing the purpose of assessment is necessary for the teachers since the purpose for which the teachers assess students determines its rationale, design, use, and interpretation of results. Popham (2014) categorizes classroom based assessment into instructional purposes (i.e., to adjust instruction to student level) and accountability purposes (i.e., to provide information to administrators). Likewise, assessment specialists classify classroom assessment purposes into two broad types: formative and summative (William 2010; Brookhart 2011). For McMillan (2014), assessment used for the formative purpose is typically associated with enhancing instruction and improving learning, whereas the summative purpose of assessment deals with summing up the learning achievements to be communicated to administrators and/or other relevant stakeholders. Furthermore, classroom assessment purposes are classified into four types recently labeled as: assessment for teaching (Care & Griffin 2009), assessment as learning (Earl 2013), assessment for learning, and assessment of learning (Lamprianou & Athanasou 2009; Popham 2014).

Griffin and Nix (1991) maintained that assessment is a broad term encompassing testing, measurement, and evaluation in the processes employed to collect information about an individual's characteristics. On the other hand, classroom assessment puts emphasis on the classroom context and excludes the term «testing» which has connotations with standardized paper and pencil tests and/or large-scale tests (Rea-Dickins 2007). Meanwhile, Lantolf and Poehner (2004) proposed the notion of dynamic assessment as the integration of assessment and instruction into a unified activity that enables learners to perform beyond their current level of functioning.

2.2. Translation quality assessment

Research on Translation Quality Assessment (TQA) indicates that scholars and professionals in translation studies encounter difficulties in evaluating a translation work (Williams 2009; Modarresi & Ghoreyshi 2018). Williams (2009) suggested that professional translators, their clients, translatological researchers, and trainee translators give more justification observing TQA. For him, TQA is a type of evaluation. In line with this assumption, Waddington (2001), who concluded that different texts must be evaluated differently, developed four different rubrics for evaluation purposes. Indeed, Translation rubrics have made progresses as a method for setting up composed rules or standards of assessments for formal, professionally-regulated article tests (Martínez-Mateo, Montero-Martínez, & Moya-Guijarro 2016).

House (2001) remarked that many factors intervene in evaluating the quality of translations that come from the analytic, comparative processes of translation criticism; however, it is the linguistic analysis that offers grounds for disputing an evalu-

ative judgment. Although language quality is a relevant factor, some professionals in interpreting studies also attach importance to strategic decisions adopted by the interpreters (Valdes & Angelelli 2003). Indeed, strategic competence, which is related to situational context, can be a main determinant of the assessment quality.

Riazi (2003) maintains that rubrics permit both translation instructors and students to survey criteria that are complex and subjective, and furthermore, give ground to self-assessment, reflection, and companion audit. Up to now, research into TQA included some standardized rubrics for evaluating the translations created by the translators (Farahzad 1992; Beeby 2000; Waddington 2001; Goff-kfouri 2004, to name but a few). Moreover, as for the assessment of translations at discoursal level, House's (2015) revised model of TQA is used to assess translation works by analyzing and comparing original and translated texts qualitatively. Kim (2009: 135) developed a model of TQA at the university level named 'meaning-oriented assessment of translations' to measure translation works quantitatively.

2.3. Interpreting quality assessment

The literature on Interpreting Quality Assessment (IQA) indicates that IQA is a timeless topic for researchers (Pöchhacker 2001), and it is different from TQA. They are different in the construct they measure whether it is translation versus interpreting skills, or it is producing written language versus oral language (Angelelli 2012). For Angelelli and Jacobson (2009: 3), «A few scholars have ventured into this new territory». Likewise, Clifford (2005) suggests that test developers identify each competency according to established theoretical frameworks, and then breaks them down into sub-traits, or sub-competencies. However, since interpreting is seen as social action (Wadensjö 1995), interpreting teachers and researchers are expected to theorize their actions while developing interpreting assessment instruments. According to Mizuno (2005: 746), «interpreters circumvent many of the difficulties by using translation strategies». Angelelli (2009), in her writings on interpreting assessment, preferred a similar approach in developing her five-point scale rubric for the American Translators Association certification exam. The sub-components she identified for determining translation quality are based on the frameworks of communication and communicative competence. Arter and McTighe (2001) concluded that grading procedures represent different levels of achievement such as superior, advanced, fair, and poor, although other types of scales can be implemented, depending on the objectives of the assessment. In this regard, Clifford (2001) suggests a similar approach to the assessment of interpreting competence, basing the particular traits to be measured on theoretical discursive frameworks related to deixis, modality, and speech acts.

Some scholars favored longitudinal studies into IQA in order to provide insights into interpreting assessment (Tiselius 2008; Han 2017). As the focus of longitudinal study in interpreting studies is mostly on interpreting assessment, Tiselius (2008) outlined the valid and reliable assessment instruments for assessing interpreting perfor-

mance (cf. Angelelli 2007; Moser 1995). The results of her study revealed that interpreters and laypeople agree on the grading of intelligibility in interpreted renditions. In his doctoral dissertation, Wu (2010) developed interpretation assessment criteria with five major components, including: 1) Presentation and Delivery, 2) Fidelity and Completeness, 3) Audience Point of View, 4) Interpreting Skills and Strategies and 5) Foundation Abilities for Interpreting, and each of these categories consists of subcategories. More recently, Han (2017) highlighted the utility of analytic rating scales such as multifaceted Rasch measurement in assessing interpreting.

As for the theoretical framework of the present study, the researchers followed the guidelines suggested by Clifford (2001) for assessing interpreting, including: (1) selection of competencies to be measured must be grounded in theory; (2) traits and their sub-components must be operationalized; and (3) assessment must be of authentic performances or as close to authentic as possible. Therefore, the competencies measured in this study are grounded in the theoretical frameworks of interpreting studies, and the validated rubric can be used as an instrument in authentic contexts. For sure, these competencies and their sub-components should not be exhaustive.

Thus, the researchers of the present study aimed to provide answers to the following three questions:

1. What factors are involved in assessing students' performance on consecutive interpreting in the Iranian context from interpreting experts and teachers' opinions?
2. Does the rubric for consecutive interpreting performance fit into a Rasch model of test performance?
3. Does the rubric for consecutive interpreting performance enjoy the psychometric properties of reliability and validity?

3. The study

3.1. Participants

To opt for the target sample, the study adopted a criterion-based selection method (LeCompete & Preissle 1993), meaning that the researchers specified some criteria for the selection of the participants. The criteria set in this phase were: a) Being an English interpreting teacher, b) Being a PhD graduate/candidate in English Literature, ELT, Linguistics, and Translation Studies, and c) Having experience of teaching interpreting courses for at least five years. During the first step, a pool of 20 interpreting experts and teachers from various parts of the country participated in this study. Six of them were interpreting experts, who worked for legal and official sectors, four of them were working for Iran's broadcasting Press TV channel, eight of them were experienced interpreting teachers teaching interpreting courses at the university level, and two of them were laymen to interpreting. Those participants who were teaching interpreting courses at the university level were from Ferdowsi University of Mash-

had, Islamic Azad University of Quchan, Imam Reza University of Mashhad, Islamic Azad University of Tehran including Science and Research Branch and North Branch, Allameh University of Tehran, and Islamic Azad University of Karaj, all located in Iran. In the second step of the study, 155 interpreting teachers were invited to respond to the interpreting evaluation checklist.

Finally, during the third step of the study, a pool of 105 BA students majoring in Translation Studies participated in the pilot study, as an integral part of developing rubric. They were selected based on availability sampling from Islamic Azad University of Quchan, Imam Reza University of Mashhad, Islamic Azad University of Tehran-North Branch, Tabaran University of Mashhad, University of Zabol, University of Kerman, University of Hamadan, University of Birjand, and University of Babol Sar. In the Iranian educational system, Translation Studies program at BA level presents three two-credit courses of interpreting for students including Interpreting (1), Interpreting (2), and Interpreting (3). The students participated in the study were selected from Interpreting (2) and Interpreting (3). They had some learning experience with consecutive interpreting performance in the classroom setting since they had already passed Interpreting (1). The students were both male and female, and they were in their semester five or six. Their language proficiency was assessed by means of Preliminary English Test (PET) including listening skills. The listening part of this test includes four parts ranging from short exchanges to longer dialogues and monologues. Having selected the students who were at the same level of language proficiency, the researchers started their research work. Out of 134 interpreting students, 105 students were made homogenous in terms of their language proficiency. The mean of the scores was 16 and standard deviation was four. Therefore, given one standard deviation above and below the mean, students whose scores obtained from PET were between 12 and 20 were selected to take part in the study (since 16-4=12 and 16+4=20).

3.2. Instrumentations

Four major instruments were used by the researchers to gather the relevant data:

The first instrument used in the study was an open-ended questionnaire designed by the researchers after reviewing several studies on consecutive interpreting, reflecting on the pre-established rubrics and criteria for interpreting assessment, and interviewing with interpreting teachers about the factors involved in assessing students' performance on consecutive interpreting (Appendix A). To seek out the beliefs of the interpreting experts and teachers regarding the factors involved in assessing students' performance on consecutive interpreting, they were interviewed face-to-face with the researchers. The second instrument was an interpreting evaluation checklist that was used to discover the important factors in measuring students' interpreting performance. The checklist included 30 items emerged from the responses obtained from the 20 interpreting experts and teachers. The checklist was administered to 155 interpreting teachers with Yes/No template.

The third instrument was the Evaluation Rubric for Consecutive Interpreting Performance (abbreviated as ERCIP) developed and validated by the researchers that can be used to measure students' performance on consecutive interpreting (Appendix B). The researchers designed and validated the relevant rubric since it was not already developed in the Iranian context. The validated rubric consisted of four factors with 22 items. Delving into the contents of the items, these new factors were named as follows: factor one: *language competence* (including 6 items), factor two: *interpreting strategies* (including 8 items), factor three: *communication ability* (including 5 items), and factor four: *personality traits* (including 3 items). The scoring procedure of the rubric was between 22 and 88.

The last instrument was a test of interpreting taken from VOA Coast to Coast News appropriate for interpreting courses at the university level. The difficulty levels of the segments were approximately the same and each segment had about five minutes long. The students were asked to listen to the monologue speech uttered by native speakers of English language and delivered the speech in Persian language sentence by sentence.

3.3. Procedure

The study followed three major steps to design, validate and apply the evaluation rubric:

During the first step of the study, the researchers gathered data from 20 interpreting experts and teachers regarding the factors involved in assessing students' performance on consecutive interpreting. The relevant data were gathered over a series of three weeks in November 2016. Each of the interviews varied in length so as for the researchers to make sure that the interviewees' responses to the questions reached saturation. The researchers made use of 'theme-based categorization' (Dörnyei 2007: 245) to categorize the responses emerged from the open-ended questionnaire. The responses were analyzed by structuring and classifying, that is, by tracing commonalities across them. Therefore, the researchers came up with the most commonly cited factors contributing to assessing students' performance on consecutive interpreting.

In the second step of the study, the researchers designed an interpreting evaluation checklist based on the common themes emerged from the interviews, including 30 items. The contents of the checklist included the factors involved in assessing students' performance on consecutive interpreting. The teachers were invited to respond to each item in yes/no template. During this step, 155 teachers who were teaching interpreting courses participated from different universities around the country from October 2016 to December 2016. Then, the data were entered in the Winsteps Rasch Measurement Software to see which items were reliable. Following this, having employed Rasch measurement, those items that fitted into the Rasch model were kept and used in developing the evaluation rubric. As the initial piloting, the researchers

asked two experts in the field of interpreting studies and two experts in testing and assessment, who held PhD in Translation Studies and TESOL and had five years of experience in teaching interpretation courses, to read the questions, evaluate them regarding the intelligibility, validity, and appropriateness, and provide the researchers with their feedback. As the results of this step, many revisions were made to the number, contents, and wordings of the items. Having received the feedback from the initial pilot group, the researchers carried out the final piloting to complete the process of construct validation for the relevant rubric.

During the third step of the study, 105 students, who were taking Interpreting (2) and Interpreting (3) courses, were asked to participate in taking the interpreting test, as the final piloting of the rubric. This step was carried out from April 2017 to July 2017. The test was indeed a performance-based task administered in the classroom setting. The researchers provided sufficient and clear instructions for the students, guiding them how to carry the task of consecutive interpreting. This was particularly crucial before the test, for very few of the subjects were expected to have tried such a task previously since this time two interpreting teachers rated them. They had already passed interpreting (1), but just their own teacher rated them. The researchers followed Giles's (2009: 168) 'effort model of consecutive interpreting' to undertake the study based on which the interpreting process was carried out in two phases: a comprehension (or listening and note-taking) phase, and a speech production (or reformulation) phase. The audio texts were played on a CD player, and participants listened to it through good quality headphones and then, they were required to render the segment speech from English language into Persian language while their interpreting performance was also recorded into a voice recorder. For the ease of scoring purposes, the students were asked to come to the class two by two in each of the target universities mentioned above by prior agreement with the classroom teachers, and the raters assessed their interpreting performance by means of the new-developed rubric.

The items were written in English language. They included one section devoted to demographic information. The typed scale comprised the items on a single page in a Likert type scale used with four levels of achievement, including superior (highest level of performance), advanced, fair, and poor (lowest level of performance). The minimum and maximum scores were 1 and 4, for each item, respectively. Then, the researchers asked two raters, who were experts in interpreting assessment, to score the students' performance on consecutive interpreting in the classroom context to ensure the inter-rater reliability of the scores. Following this, to discover the major constructs or factors of the rubric, the researchers employed Exploratory Factor Analysis (EFA), as one of the methods of construct validation. That is, EFA was implemented to check the construct validity of the newly-developed evaluation rubric for interpreting assessment performance. Finally, Structural Equation Modeling (SEM) was used to test whether measures of a construct are consistent with the researchers' understandings of the nature of that construct (or factor).

4. Data analysis and Findings

4.1. Interpreting evaluation needs

To find out the factors involved in assessing students' performance on consecutive interpreting, the researchers made use of descriptive statistics to identify the commonalities suggested by interpreting teachers and experts. Having analyzed and categorized the most common factors obtained from the open-ended questionnaire, the researchers came up with 38 factors emerged from the interviews held with the participants. The 10 most commonly-cited factors from the most to the least were as follows: 1) fluency (6.7%), 2) meaning accuracy (6%), 3) listening skills (5.6%), 4) stress and intonation (5.2%), 5) note-taking strategies (5.2%), 6) interpreting ability (4.9%), 7) background knowledge (4.5%), 8) interaction (4.5%), 9) anticipation (3.7%), and 10) knowledge of interpreting theories and models (3.7 %). Likewise, the frequency of the factors cited by the 20 interpreting experts and teachers from the most to least was as follows: 1) fluency (18 times), 2) meaning accuracy (16 times), 3) listening skills (15 times), 4) stress and intonation (14 times), 5) note-taking strategies (14 times), 6) interpreting ability (13 times), 7) background knowledge (12 times), 8) interaction (12 times), 9) anticipation (10 times), and 10) knowledge of interpreting theories and models (10 times). This indicates that the frequency of the factors, as emerged from the responses by interpreting experts and teachers, was rather high. For example, out of 20 participants in this phase of the study, nearly all of them (including 18 individuals) believed that fluency was a determining factor for consecutive assessing consecutive interpreting.

4.2. Reliability and validity of ERCIP

As for the psychometrics properties of ERCIP, the researchers opted for Rasch analysis, EFA and SEM to validate the relevant rubric. Having analyzed the factors emerged from the interviews with the experts and teachers and reflecting on the previous literature on IQA, the researchers provided an evaluation checklist in yes/no template. Then, the checklist, consisting of 30 question items, was distributed to 155 interpreting teachers teaching at the university level from different universities.

Having gathered the data from the responses to the checklist, the researchers, initially, used Rasch analysis to confirm its unidimensionality. To run Rasch, WINSTEPS (version 3.63.0) was employed. Following this, to determine the number of factors underlying the scale, using SPSS (Version 22), the researchers performed EFA and SEM. To clarify the statistical procedures utilized here, it should be mentioned that Rasch measurement reveals the major trait, it is unable to detect the fuzzy dimensions (sub-components) so that the researchers ran SEM to reveal the sub-components of the major trait. SEM was also used to confirm the results obtained via EFA.

Table 1. Person reliability and item reliability for EIRP

PERSON	75 INPUT	75 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	15.1	25.0	.49	.47	1.00	.1	.97	-.1
P.SD	2.8	.0	.55	.04	.20	1.3	.26	1.1
REAL RMSE	.47	TRUE SD	.30	SEPARATION	2.63	PERSON RELIABILITY		.79
ITEM	25 INPUT	25 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	45.2	75.0	.00	.27	.99	.3	.97	.2
P.SD	11.9	.0	.79	.03	.13	1.8	.20	2.0
REAL RMSE	.27	TRUE SD	.75	SEPARATION	2.77	ITEM RELIABILITY		.88

First, to confirm the uni-dimensionality of the scale, Rasch measurement was applied employing WINSTEPS software (Linacre 2009). The overall analysis of the items yielded an item separation index of 2.77 with an item reliability of .88, and a person separation index of 2.63 with a person reliability of .79, which indicates quite precise measurement (see Table 1).

Table 2. Item statistics and fit statistics for EIRP

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
7	36	75	.57	.24	1.31	4.9	1.38	5.0
8	37	75	.51	.24	1.24	4.0	1.32	4.3
18	50	75	-.25	.25	1.11	1.1	1.28	2.3
19	31	75	.86	.24	1.16	2.3	1.26	3.0
3	39	75	.40	.24	1.14	2.3	1.16	2.3
22	39	75	.40	.24	1.14	2.3	1.16	2.3
10	28	75	1.04	.25	1.01	.1	1.10	1.0
4	43	75	.17	.24	1.08	1.3	1.07	.9
23	43	75	.17	.24	1.08	1.3	1.07	.9
9	51	75	-.32	.26	1.06	.6	1.07	.6
13	12	75	2.24	.32	1.03	.2	.97	.0
15	36	75	.57	.24	1.02	.4	1.01	.2
16	57	75	-.74	.28	.95	-.3	.86	-.8
11	50	75	-.25	.25	.92	-.8	.90	-.9
14	53	75	-.45	.26	.92	-.7	.81	-1.4
20	54	75	-.52	.26	.91	-.7	.90	-.7
17	54	75	-.52	.26	.89	-.9	.82	-1.3
24	48	75	-.13	.25	.88	-1.5	.84	-1.6
2	54	75	-.52	.26	.87	-1.1	.80	-1.4
5	49	75	-.19	.25	.87	-1.5	.83	-1.6
12	32	75	.80	.24	.87	-2.1	.84	-2.2
21	54	75	-.52	.26	.87	-1.1	.80	-1.4
1	49	75	-.19	.25	.86	-1.6	.81	-1.8
25	65	75	-1.49	.35	.85	-.6	.67	-1.1
6	66	75	-1.62	.36	.83	-.6	.61	-1.3
MEAN	45.2	75.0	.00	.26	.99	.3	.97	.2
P.SD	11.9	.0	.79	.03	.13	1.8	.20	2.0

As displayed by table 2, 25 items fitted the Rasch model, following the criteria suggested by Bond and Fox (2007). Items which do not fit the Rasch model have infit mean square (MNSQ) indices outside the acceptable range of 0.70–1.30. Misfitting items are signs of multi-dimensionality and model deviance. 25 items were found to have an infit MNSQ index inside the acceptable boundary, as illustrated by the column

«infit MNSQ». The rest of the items were outside the acceptable boundary. Therefore, the checklist evaluation was reduced to 25 items after running Rasch analysis.

Following this step, the researchers developed the evaluation rubric including 25 items on a single page in a Likert type scale with four levels of achievement including superior (highest level of performance), advanced, fair, and poor (lowest level of performance). The minimum and maximum scores were 1 and 4, for each item, respectively. Each item in the questionnaire was designed to measure an aspect of the components of the students' consecutive interpreting performance. Having distributed the rubric to the raters to assess the students' consecutive interpreting performance, the data were entered in SPSS (Version 22) for doing further analysis by running Factor Analysis. That is, the researchers used SPSS Software to revalidate and determine the underlying constructs of the rubric by means of factor analysis.

Table 3. KMO and Bartlett's test for EIRP

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.701
Bartlett's Test of Sphericity	Approx. Chi-Square	1503.394
	df	233
	Sig.	.000

Initially, the researchers checked the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) value that ranges from 0 to 1, with .6 suggested as the minimum value for a good factor analysis. The Barlett's Test of Sphericity value should be significant (i.e. the Sig. value should be .05 or smaller) (Tabachnick & Fidell 2001). In this study, the KMO value was .701, which was acceptable, and the Bartlett's test was significant ($p=.000<.05$); therefore, factor analysis was appropriate (see Table 3).

Then, the pilot study was conducted to revalidate the rubric. The newly-made rubric, consisted of 25 items, was employed during the first phase of the pilot study. To explore the possible nature of the underlying constructs, factor analysis was run.

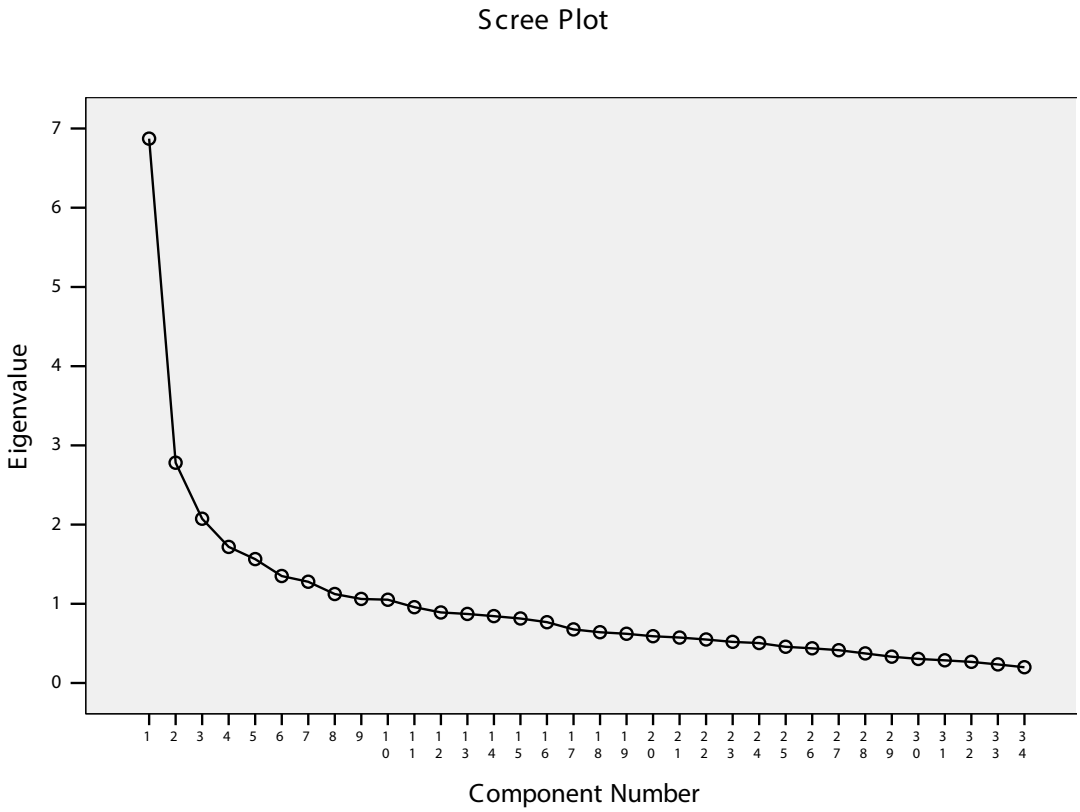
During this phase of the study, PCA extracted seven factors with eigenvalues greater than 1.0 which accounted for 77% of the variance (see Table 4). The loadings of 24 items were 0.40 or greater on any factor. Put it another way, item 25 was not found to have loadings of 0.40 or higher on any factor. Therefore, this item was removed from the rubric. The newly-developed rubric consisted of 24 items.

Following this, the researchers used the Scree Test to decide on the number of factors to retain for rotation. Given the natural bend or break point in the data where the curve flattens out, the results of the Scree Test illustrated that a four-factor solution might provide a more parsimonious grouping of the items (Figure 1).

Table 4. Factors extracted from PCA for EIRP

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.294	21.177	21.177	5.294	21.177	21.177
2	4.116	16.463	37.640	4.116	16.463	37.640
3	2.518	10.073	47.713	2.518	10.073	47.713
4	2.306	9.226	56.939	2.306	9.226	56.939
5	1.954	7.815	64.754	1.954	7.815	64.754
6	1.636	6.543	71.297	1.636	6.543	71.297
7	1.418	5.672	76.970	1.418	5.672	76.970
8	1.216	4.863	81.832			
9	.825	3.301	85.133			
10	.784	3.136	88.269			
11	.543	2.174	90.443			
12	.479	1.914	92.357			
13	.441	1.764	94.121			
14	.372	1.488	95.609			
15	.314	1.256	96.866			
16	.224	.894	97.260			
17	.212	.849	97.609			
18	.144	.575	98.285			
19	.072	.287	98.572			
20	.064	.257	98.829			
21	.039	.156	99.185			
22	.029	.115	99.403			
23	.023	.111	99.531			
24	.018	.091	99.783			
25	.012	.072	100.000			
Extraction Method: Principal Component Analysis.						

Figure 1. Scree test for EIRP



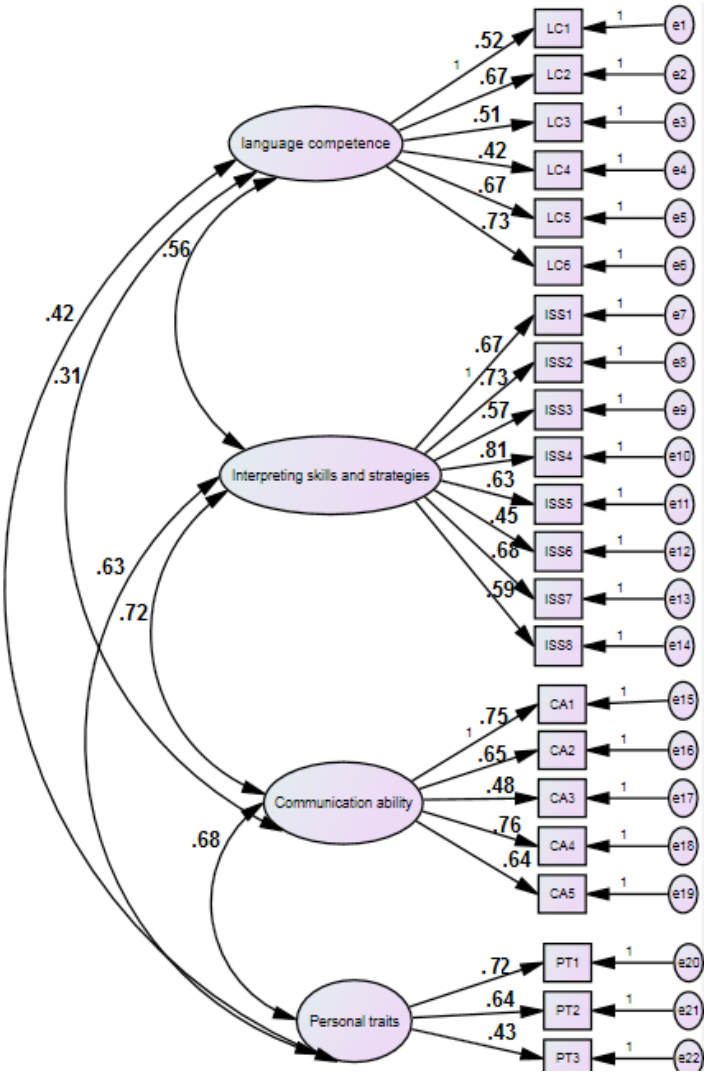
Then, oblique rotation was inspected. Varimax with Kaiser Normalization resulted in a rotated component matrix which appropriately represented the underlying factor structure, as displayed in table 5. With reference to this table, the first factor consisted of *7* items, the second factor consisted of *9* items, the third factor consisted of *5* items, and the fourth factor consisted of *3* items. The whole items following the factor rotation consisted of 24 items.

Following this, the results obtained from Amos 20 showed a good fit to the data. Since some measurement models did not show adequacy to the data, some modifications were made on the model (see figure 2). These modifications included the removal of one item from factor one, and one item from factor two due to low loadings. The goodness-of-fit of the model improved substantially following modification: V2/df was 2.34, less than the cut-off point of 3; RMSEA was .072, less than .08; and GFI, CFI, and TLI were .91, .92, and .91, respectively, all above the suggested cut-off point of .90.

Table 5. Rotated component matrix[a] for EIRP

	Component			
	1	2	3	4
Q22	-.841			
Q3	-.841			
Q23	.795			
Q4	.795			
Q8	-.602			
Q9	.602			
Q11	.560			
Q5		.858		
Q24		.834		
Q21		.780		
Q2		.780		
Q7		-.709		
Q12		.661		
Q18		.431		
Q19		.454		
Q15		.412		
Q14			.620	
Q1			.579	
Q17			.469	
Q20			.343	
Q16			.352	
Q13				.618
Q10				.426
Q6				.379
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.				
a. Rotation converged in 5 iterations.				

Figure 2. Measurement model of ERCIP



Therefore, the final rubric consisted of four factors with 22 items (see Appendix B for the validated questionnaire). Considering the contents of the items, these new factors were named as follows: factor one: *language competence* (including *6* items), factor two: *interpreting strategies* (including *8* items), factor three: *communication ability* (including *5* items), and factor four: *personality traits* (including *3* items).

5. Discussion and Conclusion

As the foremost purpose of the study was to carefully measure the reliability and validity of ERCIP, the researchers first interviewed with 20 interpreting experts and teachers to discover the factors involved in assessing students' performance on consecutive interpreting. Based on the commonalities emerged from the responses, they came up with 38 common factors. Then, the researchers designed an interpreting evaluation checklist including 30 items with reference to these common factors. Afterward, the checklist was distributed to 155 interpreting teachers around the county.

Rasch Measurement determined that 25 items of the checklist fitted the Rasch model. Following this, the researchers developed the evaluation rubric including 25 items on a single page in a Likert type scale representing four levels of achievement including superior (highest level of performance), advanced, fair, and poor (lowest level of performance). Finally, 105 students, who were taking Interpreting (2) and Interpreting (3) courses, participated in the interpreting test, as the final piloting of the rubric. Having scored each student's performance on interpreting within the classroom setting by two interpreting raters, the researchers ran EFA and SEM to validate the rubric. The validated rubric included four major factors with 22 items. Having studied the contents of the items comprising each factor, the researchers named the factors *language competency*, *interpreting strategies*, *communication ability*, and *personality traits*. The validated scale goes from 1 to 4 in which number 1 is seen as the lowest level of performance and number 4 is indicative of the highest level of performance. The results of the study are aligned with the previous research studies carried out by Angelelli (2009) who developed a rubric for assessing interpreting with an emphasis on contextualization cues and discoursal factors. However, the rubric developed here included other important factors such as contextual cues, cohesion and coherence.

Major conclusions can be drawn from the present study. The study determined the underlying constructs of consecutive interpreting assessment in the Iranian context. The constructs were then operationalized based on the identified assessment criteria using both in-depth qualitative content analysis and sophisticated statistical analyses. Williams (2013) noted that in addition to helping students to see how the task identifies with the course content, a common rubric can build students specialist in classrooms through transparency. Moreover, intercultural issues matter in specific contexts. The researchers concluded that the students who were familiar with the intercultural issues could act more effectively in the process of interpreting. Some of the students were more neutral, and were not at ease while doing interpreting. This shows that personality traits such as lack or inadequate self-efficacy and self-confidence partly account for the quality of interpreting. Furthermore, the study revealed that some students lacked the ability to make use of non-verbal communication like facial expression and gesture to emphasize what they were rendering. However, students who participated actively instead of being impartial outperformed other students. This indicates that they have developed intercultural competence, that is, they have gained competence of cultural issues that lead to better performance. For instance, knowledge of cultural differences helps interpreters and translators not to interpret or translate a message to avoid cultural conflicts (Modarresi & Moein Khakshour 2018).

Both data and findings of the present study were generated and inferred from the interpreting experts and teachers who participated in this research study. Thus, their judgments on the assessment criteria represented a valid consensus of a group of practitioners. Based on the empirical results of this study, an attempt was made to produce a working document of construct specifications for the interpreting examinations. As the items contributing to the interpreting assessment show, the task of consecutive

interpreting is hard to be captured by a few elements, and mainly the four factors for interpreting assessment in the Iranian context encompass not only linguistic and communicative abilities but also interpreting strategies and personal traits. Since the concept of factor is abstract in nature, and it refers to the underlying components of the scale, the researchers elaborate on the grouping of the items that represent each factor in more details subsequently.

The first factor, named *language competency,* consists of six items, and it mainly entails knowledge of language form and meaning. Perhaps, one of the greatest challenges facing learners in the interpreting classrooms is language proficiency (Mellinger & Jiménez 2019). Indeed, a good interpreter has knowledge of syntax, semantics and pragmatics, and he or she possesses adequate vocabulary repertoire. The essential components of language including grammar, vocabulary and pronunciation are significant in the evaluation of interpretation. However, among language skills, listening skill is documented to be more important in the evaluation, as the results of the Rasch measurement regarding the participants' responses to the evaluation checklist showed. Listening in interpreting is more complicated than general listening since the want-to-be interpreter should not only understand the message but also activate his or her memory to retain and render the message. In addition, this factor measures the extent to which interpreting students have mastery over stress and intonation. This may be due to the fact that knowledge of pronunciation features, especially suprasegmental ones, help them to distinguish the words and comprehend the sentences more easily. Lee (2013), in his doctoral dissertation on interpreting education, also pinpointed the importance of segmentation in interpreting performance.

The second factor, *interpreting strategies,* consists of eight items. This factor measures knowledge of interpreting strategies such as anticipation and also the ability to do the tasks of listening, retaining information and note taking at the same time. For instance, anticipation, which is defined as the ability to make calculated guesses about what might be said later on (Chernov 2004), is a facilitative interpreting strategy. The present study confirmed the importance of this strategy in assessing performance and its capacity to be measured. According to Riccardi (1995: 174), «it allows interpreting students to use a minimum amount of processing efforts to reduce the negative effects of cognitive constraints». Indeed, scholars may differently configure their conceptualizations of the various strategies required during the interpreting tasks, but all of these models rely on the assumption of adequate facility in at least two languages (Russo 2011).

The third factor, called *communication ability*, consists of five items. The results of the study showed that the conveyance of non-verbal communication is effective for good performance, and interpreting students should focus on both verbal and non-verbal communications. Del Pozo-Triviño and Fernandes del Pozo (2018) also highlighted the importance of training in communication for interpreters. Moreover, background knowledge that is influential in consecutive interpreting success has been another key factor. Indeed, interpreting students should acquire not only world knowledge but also necessary background knowledge of the subject.

The fourth factor, named *personality traits*, consists of three items. The contents of the items represent personality attributes including self-confidence, aptitude and ability to control breath. Self-confidence is determining in interpreting since face-to-face interaction in public settings and the live nature of interpreting require the interpreters to be self-confident. Actually, some interpreters may not be suited for interpreting jobs because of their personality traits (Riccardi 2005). Likewise, recently, some interpreting programs opt for aptitude exams to decide whether students are particularly suited to the interpreting task (Mellinger & Jiménez 2019). Thus, aptitude is also measured by the rubric because there are students who lack the natural ability for this job.

Finally, the study offers major practical implications for interpreting students and interpreting teachers. Interpreting students are suggested to take into account those factors that are important in doing consecutive interpreting tasks and work on their language proficiency and communication skills. They are suggested to work on interpreting strategies while listening to the segments and do not think of consecutive interpreting task just as a listening and testing task without resorting to different techniques proposed by scholars in the field. Interpreting teachers are also recommended to use the evaluation rubric while assessing students' interpreting performance since objective assessments based on assessment grids determine the ability of the students more precisely. They need to make students aware of the evaluation rubric based on which they are assessed in their interpreting performance. Interpreting teachers are also recommended to diagnose students' weaknesses with respect to personality traits and hold sessions to minimize their weaknesses in this regard. While measuring students' interpreting performance, they can ask their colleagues to act as the second rater so that they can check inter-rater reliability issues in assessing students' interpreting performance.

Finally, we suggest that the door is open for conducting further research concerning interpreting quality assessment in order to develop a comprehensive picture of this area in relation to interpreting performance.

6. References

- Angelelli, C. V. (2007). Assessing medical interpreters: The language and interpreting testing project. *The Translator 13* (1), 63-82.
- Angelelli, C. V. (2009). Using a rubric to assess translation ability: Defining the construct. In C. V.
- Angelelli, & H. E. Jacobson, *Testing and assessment in translation and interpreting studies,* 13-48. John Benjamins Publication Co.: Amsterdam.
- Angelelli, C. V. (2012). Testing and assessment in translation and interpreting studies. In Y. Gambier, & L. V. Doorslaer (Eds.), *Handbook of Translation Studies*, volume 3, 172-177. Amsterdam: John Benjamins Publishing Company.

- Angelelli, C. V., & Jacobson, H. E. (2009). Introduction. In C. V. Angelelli, & H. E. Jacobson, *Testing and assessment in translation and interpreting studies* (pp. 1-10). John Benjamins Publication Co.: Amsterdam.
- Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom*. Thousand Oaks, CA: Corwin Press.
- Bachman, L. F. (2014). Ongoing challenges in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment,* Volume 3, 1586-1603. Oxford: John Wiley and Sons.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests,* Volume 1. New York: Oxford University Press.
- Beeby, A. (2000). *Teaching translation from Spanish to English*. Ottawa: University of Ottawa Press.
- Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ.: Lawrence Erlbaum.
- Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, *30* (1), 3-12.
- Campbell, S., & Hale, S. (2003). Translation and interpreting assessment in the context of educational measurement. *Translation Today: Trends and Perspectives* 14, 205-224.
- Care, E. & Griffin, P. (2009). Assessment is for teaching. *Independence* 34 (2), 56-59.
- Clifford, A. (2001). Discourse theory and performance-based assessment: Two tools for professional interpreting. *Meta* 46 (2), 365-378.
- Clifford, A. (2005). *A preliminary investigation into discursive models of interpreting as a means of enhancing construct validity in interpreter certification.* Doctoral dissertation, University of Ottawa. Ottawa, Canada.
- Cohen, A. D. (1994). *Assessing language ability in the classroom*. Boston: Heinle & Heinle Publishers.
- Del Pozo-Triviño, M., & Fernandes del Pozo, D. (2018). What public-service agents think interpreters should know to work with gender violence victims. The 'Speak Out for Support' (SOS-VICS) project. *Sendebar*, 29, 9-33.
- Dornyei, Z. (2007). *The psychology of the language learner individual differences in second language acquisition.* Mahwah, NJ: Lawrence Erlbaum Associates.
- Earl, L. M. (2013). *Assessment as learning: Using classroom assessment to maximize student learning* (2nd ed.). Thousand Oaks, California: Corwin.
- Farahzad, F. (1992). Testing achievement in translation classes. In C. Dollerup, & A. Loddergard (Eds.), *Teaching translation and interpreting* (pp. 271-278). Amsterdam/ Philadelphia: John Benjamins.
- Gile, D. (1995). *Basic concepts and models for interpreter and translator training*, Volume 8. Amsterdam/Philadelphia: John Benjamins Publishing.

- Gile, D. (2009). *Conference interpreting, historical and cognitive perspectives*. London/New York: Routledge.
- Goff-Kfouri, C. A. (2004). Testing and evaluation in the translation classroom. *Translation Journal*, *8*(3), 7-13.
- Griffin, P., & Nix, P. (1991). *Educational assessment and reporting*. Sidney: Harcout Brace Javanovich Publisher.
- Han, C. (2017). Using analytic rating scales to assess English–Chinese bi-directional
- interpreting: A longitudinal Rasch analysis of scale utility and rater behaviour. *Linguistica Antverpiensia, New Series: Themes in Translation Studies, 16*, 196-215.
- Hatim, B., & Mason, I. (1997). *The translator as communicator*. London & New York: Routledge.
- House, J. (1997). *Translation quality assessment: A model revisited*. Nehren: Gunter NarrVerlag Tubingen.
- House, J. (2001). Translation quality assessment: Linguistic description versus social evaluation. *Translators' Journal* 46 (2), 243-257.
- House, J. (2015). *Translation quality assessment: Past and present*. London and New York:Routledge.
- Kim, M. (2009). Meaning-oriented translation assessment. In C. V. Angelelli & H. E.Jacobson, *Testing and assessment in translation and interpreting studies*, 123-157. Amsterdam: John Benjamins.
- Lamprianou, I. & Athanasou, J. A. (2009). *A teacher's guide to educational assessment* (Revised edition). Rotterdam: Sense.
- Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment: Bringing the past into the future. *Journal of Applied Linguistics* 1(1), 49-74.
- LeCompete, M. & Preissle, J. (1993). *Ethnography and qualitative design in education research*. San Diego, CA: Academic Press.
- Lee, S. (2013). *Ear-voice span and syntactic makeup of segments in simultaneous interpretation of Korean and English unscripted speeches*. Unpublished doctoral dissertation. Seoul: Hankuk University of Foreign Studies.
- Linacre, J. M. (2009). Winsteps (Version 3.68) [Computer Software]. Beaverton, Oregon: Winsteps.com.
- Martínez-Mateo, R., Montero-Martínez, S., & Moya-Guijarro, A. J. (2016). The modular assessment pack: A new approach to translation quality assessment at the directorate general for translation. *Perspectives*, *17*, 1-31.
- McMillan, J. H. (2014). *Classroom assessment: Principles and practice for effective standards based instruction* (6th ed.). Boston: Pearson Education.
- Mellinger, C. D., & Jiménez, L. G. (2019). Challenges and opportunities for heritage language learners in interpreting courses in the U.S. context. *Modern Languages & Literature Faculty Publications, 44*, 950-974.
- Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation* 7 (25), 1-10.

- Mizuno, A. (2005). Process model for simultaneous interpreting and working memory. *Meta: Translators' Journal*, *50* (2), 739-752.
- Modarresi, Gh., & Ghoreyshi , S. V. (2018). Student-centred corrections of translations and translation accuracy: A case of BA translation students. *Translation Studies, 15* (60), 11-28.
- Modarresi, Gh., & Khakshoor, M. (2018). Translation of cultural taboos in Hollywood movies in professional dubbing and non-professional subtitling. *Intercultural Communication Research, 47 (6), 454-473.*
- Moser, P. (1995). *Survey on expectations of users of conference interpretation*. Geneve: AIIC.Moss, A., & Holder, C. (1988). *Improving student writing*. Dubuque, IO: Kendall/ Hunt.
- Muñoz, M. R. (2012). Cognitive and psycholinguistic approaches. In M. C. Millán & F. Bartrina, *Handbook of translation studies* (pp. 241–256). London: Routledge.
- Newmark, P. (1988*). A textbook of translation*. New York: Prentice-Hall International.
- Pöchhacker, F. (2001). Quality assessment in conference and community interpreting. *Meta, 46*(2), 410–425.
- Popham, W. J. (2014). *Classroom assessment: What teachers need to know*. Boston: Pearson Education.
- Rea-Dickins, P. (2008). Classroom-based language assessment. In E. Shohamy & N. H. Hornberger (Eds.) *Encyclopedia of language and education: Language testing and assessment* (pp. 257-271). New York, NY: Springer.
- Riazi, A. M. (2003). The invisible in translation: The role of text structure. *The Translation Journal*, *7*(2), 1-8.
- Riccardi, A. (2005). On the evolution of interpreting strategies in simultaneous interpreting. *Meta, 50* (2), 753-767.
- Russo, M. (2011). Aptitude testing over the years. *Interpreting, 13*(1), 5-30.
- Sawyer, D. B. (2004). *Fundamental aspects of interpreter education: Curriculum and assessment*. Amsterdam & Philadelphia: John Benjamins.
- Stobart, G., & Gipps, C. (1997). *Assessment: A teacher's guide to the issues* (3rd ed.). London: Hodder & Stoughton.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). New York: HarperCollins.
- Tiselius, E. (2008). Exploring different methods for studying expertise. In N. Hartmann (Ed.), *Proceedings of the 49th annual conference of the American translators association* (pp. 119–148). Alexandria: ATA.
- Valdes, G., & Angelelli, C. V. (2003). Interpreters, interpreting and the study of bilingualism. *The Annual Review of Applied Linguistics 23*, 58–78.
- Waddington, C. (2001). Different methods of evaluating student translations: The questionof validity. Retrieved August 2016 from: id.erudit.org/iderudit/004583ar.
- Wadensjö, C. (1995). Dialogue interpreting and the distribution of responsibility. *Hermes, Journal of Linguistics 14,* 111-130.

- Walvood, B. E., & Anderson, V. J. (1998). *Effective grading*. San Francisco: Jossey-Bass.
- Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. London: Jossey-Bass Publishers.
- Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 18-40). New York: Routledge.
- Williams, M. (2009). Translation quality assessment. *Mutatis Mutandi, 2* (1), 3-23.
- Williams, M. (2013). A holistic-componential model for assessing translation student performance and competency. *Mutatis Mutandis: Revista Latinoamericana de Traducción*, *6*(2), 419-443.
- Wilss, W. (1998). Decision making in translation. In M. Baker (Ed.), *Routledge encyclopedia of translation studies* (pp. 57-60). London/New York: Rutledge.
- Wu, S. C. (2010). Assessing simultaneous interpreting: A study on test reliability and examiners assessment behavior (Unpublished PhD thesis). The School of Modern Languages, Newcastle University.

Appendices

Appendix A: Interview questions

Hello,

Dear colleague, first of all thank you for your time and your consideration, I am going to conduct a research regarding factors involved in assessing students' performance on consecutive interpreting. I will be thankful if you could help me to conduct this interview.

- How do you generally measure your students' performance on consecutive interpreting?
- To what extent do you think that the current evaluation rubrics for interpreting performance are adequate in measuring interpreting performance?
- What factors do you personally think are more important in measuring students' consecutive interpreting performance?
- Technically speaking, which factors are more seminal from both linguistic and cultural perspectives?

Appendix B: The Evaluation Rubric for Consecutive Interpreting Performance (ERCIP)

	Factors	Items	Poor	Fair	Advanced	Superior
1	*Language Competency*	Conveying the meaning accurately				
2		Conveying the meaning fluently				
3		Transferring stress and intonation				
4		Knowledge of genre and register				
5		Mastery over English listening skills				
6		Knowledge of English vocabulary				
7	*Interpreting strategies*	Knowledge of interpreting models				
8		Being involved while interpreting				
9		Anticipating what the speaker might say				
10		Doing tasks simultaneously				
11		Voice quality				
12		Use of note-taking techniques				
13		Using the first person while interpreting				
14		Using the third person for clarifications				
15	*Communication ability*	Attending to non-verbal communication				
16		Having background knowledge				
17		Having intercultural competence				
18		Being aware of contextual cures				
19		cohesion and coherence				
20	*Personal traits*	Having self-confidence				
21		Ability to control breath				
22		Having personal aptitude for interpreting				