AN OVERVIEW OF ASSESSMENT IN INTERPRETING – A CONVERSATION WITH OUR COLLEAGUES IN LANGUAGE TESTING

Uma visão geral da avaliação em interpretação –
Uma conversa com nossos colegas de medidas de avaliação em língua estrangeira

DOI: 10.14393/LL63-v35n2-2019-8

Reynaldo J Pagura*

ABSTRACT: This article suggests that the interpreter education community would benefit from cross-fertilization with the language testing community, which has been around much longer. The main principles used in testing speaking (or communicative competence) developed along the last decades by language testing experts and institutions can – and should – be applied to the testing of interpreter performance, mainly in high stakes examinations, such as those used for certification, employment at an international institution, or graduation from an educational program. The concept and some relevant studies concerning the notion of quality in professional interpretation are also discussed and suggested as ancillary help to build the constructs used in said examinations.

KEYWORDS: interpreting assessment. Validity. Reliability. Test constructs. Quality in interpreting.

RESUMO: O presente artigo sugere que a comunidade de educadores de intérpretes pode se beneficiar de uma colaboração com a comunidade especializada em medidas de avaliação em língua estrangeira, existente há mais tempo. Os princípios básicos adotados para a avaliação da habilidade da fala (ou competência comunicativa), desenvolvidos ao longo das últimas décadas, por especialistas e instituições envolvidas com medidas de avaliação em língua estrangeira, podem – e devem – ser adotadas para a avaliação de desempenho de intérpretes, principalmente em exames de alta importância, como os utilizados para a certificação, emprego em instituições internacionais ou formatura de programas educacionais. O conceito e alguns estudos relevantes a respeito da noção de qualidade em interpretação também são discutidos e sugere-se que possam auxiliar na elaboração dos construtos utilizados nas mencionadas situações avaliativas.

PALAVRAS-CHAVE: Avaliação de interpretação. Validade. Confiabilidade. Construtos de provas. Qualidade em interpretação.

* PhD, University of Illinois at Urbana-Champaign. ORCID: 0000-0003-3039-0516. E-mail: rpagura(AT)illinois.edu.

1 Introduction: Defining the scope

This article looks at assessing interpreting from two overlapping points of view – interpreter education and interpreter certification. It will focus not only on conference interpreting but also on community interpreting, an "umbrella" label in itself that includes, depending on who uses the term, legal/court interpreting, health care interpreting, educational interpreting, and religious interpreting, among other possibilities.

According to Pöchhaker (2001, p. 411),

> there is something to gain by taking a comprehensive, unifying view on interpreting before focusing on a particular domain for specific investigations, I will define 'interpreting' as a conceptual spectrum of different (proto)types of activity. ... 'conference interpreting' and 'community interpreting' are understood not in terms of a dichotomy but as different areas along a spectrum which ranges from interpreting in an international sphere of interaction, among representatives of entities based in different 'national' or multi-national environments, to interpreting within an institution of a particular society or social community, between individuals representative of that institution.

Also, it discusses the attempts to define what quality interpretation in the professional field really is and how difficult it has been to validate it. As a last introductory remark, it should be made clear that the term "interpreting" in this article is used to refer to spoken language interpreting, as opposed to sign language interpreting, which is not part of the author's expertise and not being discussed here.

2 Why "a conversation with our colleagues in language testing"?

While Interpreting Studies literature is no longer in its infancy, as it was some decades ago, issues of assessment have been treated from a more scientific perspective only relatively recently in the field – just as in Translation Studies. Assessment in foreign language teaching, mainly in English as a Foreign Language, has been part of the educational panorama for over a century. The first CPE (Certificate of Proficiency in English) examination, by the University of Cambridge, was offered in 1913 (WEIR, 2012). TOEFL (by the American testing organization

ETS, the largest testing organization in the world) was first offered in 1964 (VIDAKOVIC; GALACZI, 2012), and its main competitor nowadays, IELTS, was first administered in 1989 (www.manhattanreview.com – accessed in November 2019). All these examinations have changed along the years and adapted to new concepts and views pervading second language teaching.

In Interpreting Studies, this "conversation" across disciplines has been called for by a few authors (MOSER-MERCER, 1994; SAWYER, 2004; ANGELELLI, 2009; SETTON; DAWRANT, 2016). It has really been tackled in a lengthier format in regard to interpreting by Sawyer and also by Setton and Dawrant (2016)

The field of (English) language teaching has, along the years, borrowed heavily from psychological tests (psychometrics) and widely adopted its terminology and methods. Some terms should be defined up front here, since they are "buzzwords" of educational assessment and used constantly (but not necessarily always accurately) in articles and discussions regarding assessment in Translation and Interpreting Studies. These first terms are *validity* and *reliability*, often appearing together, since they depend on each other. In very simple and generic terms, *validity* implies that a test is valid "if it does what it is intended to do, which is typically to act as an indicator of an abstract concept … which it claims to measure" (DAVIES et al., 1999/2002, p. 221). A reliable test is one whose results remain the same, regardless of where it is given and, very important, who grades it. The concept of reliability is, undoubtedly, the hardest to achieve in testing interpreting, and will be at the core of our discussion in this article. Multiple-choice tests, for instance, have a very high level of reliability since they can be computer graded and only one answer is possible. This issue has been so important in testing, mainly in the United States, that the TOEFL, for many decades after its implementation, would not have a measure of spoken ability, and claimed that writing ability could be measured by multiple-choice items. If at all possible, the multiple-choice item type is used in all high-stakes tests, which can be more subject to legal issues, since it is very reliable, not depending on individual judgment.

Another relevant term to define early on is *construct*. "A construct can be defined as an ability or set of abilities that will be reflected in test performance, and about which inferences can be made on the basis of test scores" (DAVIES et al., 1999/2002, p. 31). A real example to help the reader grasp how *construct* interacts with *validity* and *reliability,* familiar to this author, but whose identification data will remain unmentioned, is the following: in a course teaching consecutive interpreting from B to A (the foreign language into the native language of the students), the instructor decided to use, as the final examination, a translation of a written passage from the foreign language into the native language. Although the language direction was the same as the practice in consecutive practiced along the semester, the written translation test was clearly an invalid construct to measure consecutive interpreting *traits* in the students. Since it was not valid, it was not a reliable measure of competence in consecutive interpretation, since those traits cannot be measured in a translation of a written passage. A construct is said to measure a *trait* or traits in a person. By assessing the performance of a person in a given test, we can make inferences of the person's ability or behavior in relation to a given trait or traits (DAVIES et al., 1999/2002).

In measures of English as a Foreign Language, the main constructs measured are usually Reading, Writing, Listening, and Speaking. We could say that in Interpreting, the constructs usually measured are consecutive interpreting and simultaneous interpreting. The consecutive construct can be subdivided in monologue and dialogue interpreting. The simultaneous construct could also be subdivided in simultaneous without text and simultaneous with text. Some interpreting examinations will also measure sight translation, as a separate construct. The description of the construct should be defined by subject specialists before the test is put together and will include items that measure the traits that allow users of the test results to verify that the student/candidate can perform the traits measured in the construct in a reliable way in all situations.

3 From Language Testing to Interpreting Testing

Considering the basic constructs tested in second language measures, as mentioned above, it is not difficult to see that the issues in interpreter testing will be more related to those in tests of speaking ability than to any other construct used in language testing, just as translation tests can draw on issues related to the test of writing. One might argue that interpreting also involves listening, just as translation involves reading. But when considering the issues relating to test format, feasibility and, above all, grading, there should remain no doubts that testing interpreting constructs will face the same difficulties as those in speaking tests. The listening and readings constructs can be easily measured (and usually are) by multiple-choice or other discrete-point types of items. Writing and speaking are not measured by use of *discrete-point* items, but are always, to keep using measurement terminology, *integrative* types of tests – a "test in which learners are required to combine various skills in answering test items" (DAVIES et al., 1999/2002, p. 83).

Assessment in interpreter education programs, occur, in a summative/predictive way, in usually three moments when high-stakes tests are given: entrance examination, mid-course examination, and final examination, although these labels will, of course, vary from one institution to another. "Conference interpreting is a testing-intensive profession," state Setton and Dawrant (2016, p. 373.) It is obvious that, like in all educational situations, there is continuous assessment of the formative type in the courses offered in a program. The format and timing will vary immensely, as it can easily be inferred.

Of these three types of examinations, the one that seems to have attracted the attention of most researchers is the entrance examination, a predictive examination (hopefully), which should define whether a candidate has aptitude or not for interpreting. In this respect see, among many others: AIIC, 1965; Keiser, 1978; Longley, 1978; Dodds, 1990; Moser-Mercer, 1994; Arjona-Tseng, 1994; Mackintosh, 1995; Seleskovitch and Lederer, 2002; Sikorski, 2010. In these and other similar publications, the authors are concerned mostly with what kinds of test items an aptitude test should include. Space constraints here do not allow it to describe each one individually. Of those mentioned, only Arjona-Tseng (1994) is concerned with validity and reliability issues and how to attain them. In most institutions, these tests are "graded" in a rather intuitive format, with members of the evaluation panel "knowing" the level of language and other traits that can predict success in an interpreter teaching program. This "method", of course, has its critics, such as Gile (2001, p. 390), who states

> However, at present, there does not seem to be a reliable evaluation system for student potential, at least judging by the fact that among the many articles that deal with the issue, none provides corroborating results. [1]

Discussions on the midcourse exam, a summative and predictive form of exam at the same time, and which is part of some conference interpreting programs, does not seem to be the subject of much research. It is mentioned in Donovan (2008) and Hewetson (2008), both as part of the same *Seminar on Assessment in Interpreting*, held at the University of Graz, in April 2008, as part of the European Masters in Conference Interpreting consortium of universities. The subject is also discussed by Setton and Dawrant (2016) with the same focus: in most cases, these exams decide if a student can proceed in the program, moving from consecutive to simultaneous interpretation, or whether they should repeat the training in consecutive or find another future profession. In many countries, however, students cannot be legally told to leave the course in the middle as was the custom in France, and all that can be done is "advise" the student to change to a different program. Sawyer (2004, p. 112) gives this kind of test two paragraphs in his 306-page book, but he raises a very important point:

[1] Original quotation: "Cependant, a l'huere actuelle, on ne semble pas disposer d'un système fiable pour le potentiel des étudiants, du moins si on en juge par le fait que parmi les nombreux articles qui traitent la question, aucun ne fait état de résultats probants." (Author's translation into English.)

> As intermediate testing has the purpose of assessing whether the candidate has the potential to continue and successfully complete the degree program, the predictive validity of this type of assessment should, by definition, be high.

The most important examination given by most (mainly conference) interpreting programs is the final examination or diploma examination, as called by certain institutions. The underlying construct here is to verify that graduating students are basically at the level of professionals, albeit at a beginning stage. Most institutions give a lot of importance to these exams, which are summative in nature, considering they bring together everything that was practiced in the program. There are cases in which a diploma is refused to students who fail those examinations – again, legal concerns are making this less common. Setton and Dawrant (2016) dedicate many pages to the issue of final examinations. This, however, seems to be a quite euro-centric discussion topic and centered on conference interpreting programs. Although undeniably very important from the educational point of view, summing up what students have acquired along their education, the reality is that in order to be hired as either staff or free-lance interpreters by international organizations, such as the U.N., the European Commission, the European Parliament, among others, candidates have to pass these organizations' own entrance examinations – regardless of having an institutional diploma or not, being a member of an association or not. In the commercial market, which provides (conference) interpreters with the bulk of their work all over the world, recruiters would rarely or never bother to ask to see any diploma, since interpreting is not a recognized profession all over the world, different from law or medicine, for instance, and therefore requiring no formal credentials from those offering their services in the field.

The same apply to more formal levels of community interpreting, mainly in the case of court interpreting. Each country – and in the United States each state and also the Federal Courts (more on this further ahead) – has its own criteria of certifying court interpreters, regardless of which training program they have completed (or not). Health care interpreters, each time more, are also subject to certification examinations by different associations or institutions, in different countries and in different states of the United States. Hlavac (2015) describes community interpreter certification in some countries (United Kingdom, Australia, Canada, and Denmark) and, very importantly, brings to the fore the creation of ISO Guidelines for community interpreting, which were released in 2014 and, as many ISO standards, should eventually be adopted by many institutions and countries.

The situation for community interpreting in health-care settings is much more haphazard, at least in the United States. Although there are certification processes already in place by associations, such as CHIA (California Healthcare Interpreting Association), the situation is far from organized as it is for court interpreting, where 43 of the 50 states adopt the certification process offered by the Consortium for State Court Certification, with minor variations, for several languages (instead of the Spanish only certification offered for Federal Courts.) Back to health-care interpreting, not only the certification by CHIA is not necessarily legally required, but hospitals and health-care providers have the liberty of making their own decisions as regarding who they call to interpret and how they test these prospective interpreters, when they do. It is not unusual to use family members or to call hospital administrative or maintenance staff who happen to speak the language needed at the moment. Jacobson (2009) describes the alarming situation of many health-care providers designing tests themselves to be used in their institution. Such tests usually only require the translation of words and phrases in isolation, and many do not even accept a paraphrase of explanation, clearly reflecting the layman conception that interpreting (and translation as a whole) is nothing more than a word or phrase substitution from one language into another and that there is a perfect equivalence of words and phrases between languages, which she labels the "conduit model." What the author proposes is that other linguistics principles should also be taken into consideration when testing health care interpreters. She suggests using Interactional Analysis and Conversation Analysis technics to include other features in such tests and goes on to suggest two rubrics: one for "Contextualization Cues" and one for "Discourse Management". Although still grounded on Linguistics only (basically Pragmatics), what she proposes is, obviously, a progress in relation to the word/phrase translation test designed by health care providers themselves. Unfortunately, although mentioning the testing buzzwords "reliability" and "validity", she never mentions how these should be achieved when testing the candidates and the tests themselves, for that matter. She does not discuss the issue of raters, let alone inter- and intra-rater reliability, which are fundamental when using rubrics for holistic assessment. Both rubrics she proposes include four levels: Superior, Advanced, Fair, and Poor, with descriptors for each level. Just a quick example shows how complicated these rubrics may become when the issue of inter- and intra-reliability is not taken into consideration. The difference between a "Superior" and an "Advanced" level classification in the Contextualization Clues, for example, lies mainly in the difference between the understanding of the difference between "demonstrates *superior* ability in understanding meaning of contextualization clues..." and "demonstrates *advanced* ability in understanding meaning of contextualization clues..." (my italics.) So, unless the issue of the use of more than one rather is included and the inter-and intra-rater reliability is taken into consideration, the situation will not improve much in relation to the word/phrase translation tests she discusses.

Little by little, community interpreting becomes more professionalized and, with this, more gate-keeping procedures are put in place – but none seeming to require a diploma or a specific certification of legal validity. How all these factors will eventually alter the importance of very formal final examinations in interpreter training programs is for us to guess at this point. And how to make interpreting and translating recognized and officially certified professions is a whole issue that has no place in this discussion here.

4 How can language assessment contribute to the validity and reliability of interpreting assessment?

As mentioned above, the speaking construct is the one most likely to contribute ideas to interpreting assessment, considering that interpreting as an assessment construct is integrative and must be tested orally. No one would dare claim that it can be tested by means of discrete-point, multiple-choice items, in a paper and pencil (or computer) test, which would have high levels of reliability, but no construct validity. This is, obviously, an issue that has been faced by the language testing community for many years – how to achieve good levels of validity and reliability in tests of speaking or, as sometimes they are called, of communicative competence. It has been an exercise of reaching a mid-point, striking a balance between validity and reliability. As quoted by Stevenson (1981), "(w)e are well aware in language testing that 'all theoretical problems ... are likely to be present in a concentrated form when trying to measure performance in a spoken language' (PERREN, 1968, p. 108)." Vidakovic and Galaczi (2012, p. 257) reinforce that idea when they say that

> Most language testing professionals would agree that the testing of speaking is not an easy endeavour, since oral assessment brings with it an array of issues which do not fit easily in a dominant psychometric paradigm.

They also remind us, mentioning a comparison used previously by Spolsky, one of the best-known language testing experts that says that measuring performance assessment is similar to a sports competition, where some disciplines can be measured with complete accuracy, such as how many meters or how high did an athlete go, while some other require human "subjective judgment of expert judges, as in gymnastics ... for example" (VIDAKOVIC; GALACZI, 2012, p. 257). The measurement of traits making up the construct of different tests of interpreter performance certainly are of the "gymnastics" type in that human experts will have to be the judges of that performance, in that, just like speaking tests, interpretation tests "involve uncontrollable variability" Vidakovic and Galaczi (2012, p. 262).

The consensus today seem to be that tests of speaking ability (or "communicative competence") are validated by the combination of two requirements: (1) a scale/rubric describing different expected levels of performance and assigning a number or a letter to each of them; and (2) that each test is graded by more than one assessor separately. Let's discuss those two requirements and see how they can be (and are, in some cases) applied to interpreter performance tests, mainly for certification purposes, using the example of the English<>Spanish Federal Court Interpreter Certification Examination, which certifies Spanish court interpreters for Federal Courts in the United States. The first requirement – a scale or rubric – is quite familiar to most language instructors as well as to most interpreting instructors. Most institutions and/or instructors have developed their own rubrics for consecutive interpreting, simultaneous interpreting, interpreting into A, interpreting into B, etc. Ideally, mainly for certification purposes, this rubric should be developed by a group of subject specialists, in discussion and consensus. The Examinee Handbook – United States Courts – Federal Court Interpreter Certification Examination, published by the Administrative Office of the United States Courts (2019, revised edition) clearly lists the names of all experts involved in the development of their English <>Spanish Federal Court Interpreter Certification Examination. It is interesting to observe that it includes court interpreters, language testing specialists, linguists, and also legal specialists, such as federal judges and attorneys. The inclusion of legal specialists is a plus for this kind of exam due to their familiarity with the language and the situations encountered in courts – something that can be out of the level of expertise of testing specialists and linguists. The second requirement – a panel of expert graders – is also exemplified by the same examination, and follow procedures that have been common in high stakes language tests for several years, not only for speaking tests, but also for writing tests. The concept behind the same test being assessed by more than one grader is to reach statistical correlation between the raters, a process called inter-rater reliability, which is fundamental to validate tests in which subjective judgment is needed for scoring. Two grades can be present at a live interview (as in the IELTS, for instance) or the exam may be proctored by one person (who does not rate) and recorded for subsequent scoring by two or more assessors. These assessors score separately and then compare their scores. If the distance between them is greater than a pre-established value (depending on the scale being used), the test is usually scored by a different grader – a third grader, when only two were initially involved, or a supervisor or coordinator, when two were already involved in the initial grading, assuming that this supervisor/coordinator is more experienced than the initial graders. This recording process with subsequent grading by two trained graders is the model adopted by the Federal Court Examination mentioned above.

In this kind of process, the pre-training of graders is of fundamental importance. Graders are trained grading previously recorded exams – another advantaged of recording – until they reach an **inter-rater reliability** of more than 0.9 (number may vary between 0 and 1 for correlation) in the practice exams graded. In more simple terms, they should give the same score to at least 90% of the practice exams graded (this number may vary from one institution to another). Also, part of this training aims at **intra-rater reliability** – that is, the same rater will assign the same score to the same practice exam graded usually at least two weeks apart. These two different measures of reliability aim at reaching coherence not only between different graders but within the same grader on different days. In serious processes, only graders reaching a pre-established inter and intra-rater reliability in the training process will assess exams of real candidates.

It is interesting to notice that the Federal Court Examination already mentioned has added another layer of reliability to their process. For each **form** (a complete test) of the examination offered (and there are several), the organizing committee establishes 220 Scoring Units, classified into three general categories and nine specific types. These categories are Grammar and Usage (grammar/verbs and false cognates/interference/literalism), General Lexical Range (general vocabulary, legal terms and phrases, and idioms/sayings) and Conservation (register and slang/colloquialisms, numbers/names, modifiers/intensifiers/emphases/interjections and embeddings/positions) – see the referred handbook for individual examples.) These are distributed in two sight translations (into English and into Spanish), one simultaneous interpretation section of monologue speech, one consecutive section, and one simultaneous section of questions and answers as found in a witness testimony. This inclusion of pre-established grading categories, clearly language-related, was severely criticized by Setton and Dawrant (2016, p. 406-409.) They claim that, “in terms of construct validity ... (it) fails to assess fidelity, the primary determinant of usefulness and quality of an interpretation" (p. 406.) But is that really so? Considering there are 220 *carefully chosen scoring units by specialists* (my italics) if a candidate misses most of those, lack of fidelity would clearly stand out. Also, getting 80% of those correct (the minimum passing score in the exam) seems to assure a quite reasonable level of equivalence between original and interpreted discourse. There are other criticisms by Setton and Dawrant (2016) but the impression is that they are too centered on conference interpreting, when the establishment of such "scoring units" would be much more complicated during the countless different situations in which conference interpreters work, if contrasted with court interpreters. Also, the Examinee Handbook mentioned above, clearly states that

> ten percent of the tests whose original score was around the cut score ... are chosen for re-scoring by a second team of raters. Raters also complete a structured holistic evaluation to supplement the objective scoring procedure. This holistic evaluation assesses the strengths and weaknesses of the candidate's overall performance ... In rare cases, the holistic evaluation may also promote a candidate with an objective score that is below but very near the pass point into the "pass" category. (p. 36)

It is clear, from the above quote from the Examinee Handbook that the committee members are aware of the possible validity problems derived from the so-called "objective scoring procedure" – so much so that variations are predicted as needed, as explained above. Also, it seems clear that there was a sort of "give and take" when the examination was put together, considering not only the somewhat limited situations in which court interpreters do their job but also the possibility of legal challenges by candidates who fail the certification examination. For a very high stakes examination, it seems that the Federal Court Interpreter Certification Examination of the U.S. Federal Courts has done the very best that could have been done for this specific situation.

5 What is quality interpretation?

The idea of what constitutes good interpretation has been debated for decades and a consensus has never been reached. It is obvious that certification examinations intend to identify professionals able to deliver “good interpretation”. The same can be said of final examinations in interpreter training programs. But what constitutes a good sample of interpretation in the professional field? This has been the subject of discussion among scholars for decades, and there have been attempts to define the notion of quality in interpreting, not only in educational settings, but also in the opinion of the final user of the services provided by interpreters. This seems to have proved more difficult than the validation of educational or certification examinations.

In the conference interpreting setting, many are the studies that resort to the use of questionnaires to be replied by delegates, that is, attendees at international events in which (mostly) simultaneous interpretation between two or more languages is provided and who have used the service of the interpreters. Among those studies, are Gile (1990), Moser (1995) Kurz (1989 and 2001), and many others. The type of questionnaires used, the sequence of the questions, and terminology used will vary immensely from one study to another. Concepts that seem to be present in all of them are those of fidelity/accuracy, language acceptance, correction and coherence, good voice and intonation, use of appropriate terminology, overall weakness of the performance, etc. Due to this immense variability and due to the fact that respondents will have very different understanding of items such as “fidelity”, “good voice”, “language coherence” and other terms used, it is almost impossible to generalize the answers deriving from these questionnaires. In a seminal article commenting on this kind of questionnaire-type studies, MacDonald (2013) brings up a series of relevant comments, relying on marketing/advertising theory specifically and the social sciences as a whole. He reports that in 1996, AIIC explored the possibility of applying for ISO quality certification and invited a specialist to analyze the issue. This specialist concluded that “quality would be difficult to control in conference interpreting precisely because it depends on so many factors beyond the interpreter" (LUCCARELLI; GREE, 2007 *apud* MACDONALD, 2013, p. 40.) Also, a key question in all those questionnaires is on "fidelity/accuracy" in relation to the original. Considering that delegates are using the services of an interpreter, it is safe to assume that they do not understand the original language in which a speech is being given, or have such a limited understanding of it that they prefer to resort to the use of the interpretation provided and listen to it in their native language. How would they be able to evaluate how faithful or how accurate the interpretation is, not even considering the different understand people have of faithfulness or fidelity in interpreting? MacDonald (2013) also raises the point that the place where a question is asked in a questionnaire (in the beginning or toward the end) or even the moment in which a questionnaire is answered will alter the results, according to marketing theory. He mentions something that all seasoned conference interpreters are well aware of: "interpreters must seek to convey the feeling that they are trustworthy by producing a discourse that 'sounds' logical and inspire confidence" (MACDONALD, 2013, p. 48). This is usually done by using the jargon that listeners expect and by "mediating" culture differences, even when omitting several details. That will rarely be perceived by the listeners, who cannot follow the original speech in a language they do not speak. "Trust is the key word... participants often choose to listen to the interpretation even if they can 'get by' without. They do so for reasons of convenience, but only as long as the interpreters inspires confidence" (DONOVAN, 2002, p. 4 *apud* MACDONALD, 2013, p. 48.) He also raises several other important issues, which we unfortunately lack the space to comment in more detail.

Still on the issue of quality in simultaneous interpreting, a panel convened by the late Miriam Shlesinger during the *International Conference on Interpreting: What do We Know and How?*, held in Turku (Finland), in August 1994, and in which several luminaries in Interpreting Studies (other than Shlesinger herself, also Karla Déjean le Féal, Ingrid Kurs, Franz Pöchhacker, Maurizio Viezzi, among others) took part, Shlesinger (1997) begins her report with the following sentence: "Quality is an elusive concept, if ever there was one" (p. 122). She also states that "AIIC is still groping both to define it and especially to devise ways of maintaining it"

(p. 122). Maurizio Viezzi (*apud* SHLESINGER1997, p. 127) ratifies what was just mentioned above:

> ... the listener is lacking one the most crucial means of assessing quality: an understanding of the source language. Thus, for example, smooth delivery may create the false impression of high quality when much of the message may in fact be distorted or even missing.

Going back to the criticisms made by Setton and Dawrant (2016, p. 407) on the assessment criteria in the Federal Court Interpreter Certification Examination, as mentioned above in a previous section of this paper, the authors point out (or complain) that “a test-taker cannot fail the exam, for example, due to poor pronunciation and intonation, strong accent, hesitation, inaudible voice, etc.” Although never explained by the organizers of the certification exam, perhaps what they are trying to prevent is just what Viezzi mentions above: “that the message may be distorted or even missing” and still sound acceptable because of a “smooth delivery”. While in conference interpreting this might be acceptable in most instances, provided that the whole sense of the speech is carried out, it may undoubtedly have catastrophic consequences in a court situation, in favor of or against a defendant, of a nature never existing in an international conference.

In modalities of community interpreting other than court/legal interpreting, fewer studies can be found, which makes sense, because this kind of interpreting is, to a certain extent, still in the process of becoming more recognized as a profession and, as a consequence, has drawn the attention of fewer scholars, even if we consider the pioneer efforts by a series of conferences tittle *The Critical Link*, which started in Canada by the efforts of Roda Roberts and others, and have met several times. Pöchhacker (2001) mentions a few attempts to define what good community interpreting is. He first mentions a study carried out in Australia in 1981, which surveyed 65 community interpreters (and not interpretation users). We can quickly see the different nature of the work being done and the professionals involved in it, if compared to conference and court interpreters. Among the many ideas of a good interpreter, responses include “knowledge of both languages and the migrant culture ... honesty, politeness and humility” (HEARN et al., 1981, p. 61 *apud* Pöchhacker, 2001). Pöchhacker (2001, p. 415) also mentions that difference between the user of interpreting, comparing conference delegates or representatives and users of community interpreting:

> Where the primary interacting parties [in community interpreting] will usually take alternating turns at speaking and listening, they are essentially different in their status as ‘representatives’ as opposed to ‘clients’ of an institution or public service. It is thus common to refer to ‘service providers’ or ‘professionals’ on the one hand and ‘non-(majority-language)-speaking clients’ on the other.[2]

The same author reports a study carried out in Montreal, Canada, involving 66 clients (speaking 11 different languages) and 288 health care workers from 30 different institutions. The highest rated qualities of a good interpreter, as replied by the health care workers, are the following, in this order: “fully understands client’s language”, “ensures confidentiality”, “points out client’s lack of understanding”, “refrains from judgement” and “translates faithfully” (MESA, 1997 *apud* PÖCHHACKER, 2001, p. 417). Pöchhaker (2001) comments that it is surprising that the item “explains cultural values” was not among the top qualities chosen (61% considered it very important) and even fewer respondents (47%) included “receive cultural explanations from the interpreter after the mediated exchange” (p. 415). One might just suppose that these two “qualities” could be understood as “points out client’s lack of understanding”, chosen by 92% of responds as very important. But the article does not go into these considerations and no further information is available in the article on the study discussed. In a study carried by himself (PÖCHHAKER, 2000, and reported in PÖCHHAKER, 2001), he “collected responses from 629 health care and social workers on interpreter qualifications and role definitions” in a study carried out in Vienna. Of the ten items proposed, only two were considered “very important”: “strictly neutral behavior” and “discreteness and

[2] Usually referred to as LEPs, in the United States, which stands for Limited English Proficiency speakers.

confidentiality", a result that would be quite different, one may infer from all other studies, if done among users of conference interpreting. It is also pointed out that 62% of the respondents "expected interpreters to explain cultural references and meanings and to formulate autonomous utterances when asked to do so by the provider" (p. 415-416.)

These features would be quite unexpected not only in the conference setting, but mainly in courtroom scenarios. Actually not explaining or making their own utterances is what is expected of court interpreters. Based on the author's own experience in court interpreting, this is actually feared by judges and other legal professionals, who always insist that we "translate just the words said", in an assumption that there is always a word-for-word correspondence between two languages. The reason for this is that they are mainly concerned with any possible deviations and alterations of a testimony by a witness or a defendant, for example and their possible consequences in the final judgment and in court records. Not being interpreting or even language experts, they fall back on the usual layman understanding of the translation process, assuming that there are perfect equivalences for words across languages. This is not usually a problem, since interpreters are not often asked to explain or justify themselves. However, most are careful enough not to give the impression that they are adding to what has been said by the original speaker. If a situation arises when the interpreter has to add something to make the idea clear, it is common to raise one's hand and use the multipurpose expression "Excuse me, your Honor" and let the judge know what one has to do.

6 Concluding Remarks

As we have just seen, there is no statistical, empirical agreement to what quality in interpretation is. However, all the multiple studies carried out can easily subsidize the construction of rubrics for use in interpreter education programs of the most varied types, as well as for use in certification exams. The situation is not different from written translation, in which the ideas of fidelity and quality have been discussed for centuries, as is widely known.

The use of rubrics or descriptors is very important in education and certification. These should be designed by interpreting specialists and, when possible, with the assistance of language testing experts, mainly when they are to be used in high stakes examinations, such as certification, hiring by an international institution (such as the U.N. or the European Commission), or graduation from an education program. Rubrics can – and should – also be used in class for formative and summative examinations in a given course, and can also be helpful for students doing self- or peer-evaluation in reflective teaching situations, even more important in multilingual classes, when instructors do not have all the language combinations of their students.

While rubrics to be used in class are of a simpler nature, those used in high stakes examinations should be carefully built to include the traits included in the construct being measured. And even more important, should be used by a panel of graders who have received previous training in their use and who have practiced before the grading activity and been found to having scored appropriate inter- and intra-rater reliability, the lack of which will result in intuitive assessments in which, as several studies show, there is a high level of disagreement among different raters. MacDonald (2013) reports on a study carried out by Peter Mead on the issue, involving interpreting instructors in Italian and Austrian universities, who were asked to assess the interpretations by five students:

> Lacking of consistency between the various assessments indicates considerable variability in standards and priorities from one assessor to another. It was emblematic, for example, that there was unanimity about awarding a pass or a fail for only three out of ten interpretations.
> Another interesting was that almost none of the seven assessors could generally be identified as a consistently higher (or lower) marker than others (MEAD, 2005 *apud* MACDONALDS, 2013, p. 40)

Wu (2010) also points out a similar situation. Quoting Campbell and Hale (2003), Wu states that “they found that ‘there exist a number of knowledge gaps’ and that evaluation is mainly ‘intuitive’ and then adds that such “gaps mainly concern the reliability and validity issues especially within the context of educational measurement” (p. 301.) The author also adds what many other studies show: that the assessment, as well as administration of interpreting exams, relies on the “experiences (sic) of individual trainers”.

It is high time the interpreter education and certification community adopted more reliable ways of assessing students, mainly in examinations that can define their future. It is not bad in itself for interpreting instructors to rely on their experience and intuition in the courses taught in an educational program – quite the opposite! The comments by experienced instructors have always helped students develop and progress until they reach the level of quality – discussed as it may be – expected of professional interpreters and the intuition of their instructors cannot do them any harm and can definitely help them get there, as long as they do what is expected of them to acquire interpreting competence – however it may be defined. However, in key assessment situations, as already mentioned, students and/or candidates to a position or a certification have the right to be assessed by means of tests constructed in such a way that they are valid and graded in a reliable manner, by a well-trained panel of raters, whose grading reliability between each other and within themselves can be claimed to be reliable enough to be fair to the profession.

References

ADMINISTRATIVE OFFICE OF THE UNITED STATES COURTS. **Examinee Handbook – United States Courts – Federal Court Interpreter Certification Examination**, Revised edition. Washington, D.C.: Court Services Office, 2019. Available at https://www.prometric.com/test-takers/search/aousc. Accessed in October 2019.

AIIC (Association Internationale des Interprètes de Conférence). **Colloque Sur L’enseignment de L’interprétation**. Paris: AIIC, 1965.

ANGELELLI, C. Using a Rubric to Assess Translation Ability: Defining the Construct. *In*: ANGELELLI, C.; JACOBSON, H. E. **Testing and Assessment in Translation and Interpreting Studies**. ATA Scholarly Monograph Series XIV. Amsterdam/Philadelphia: John Benjamins, 2009. https://doi.org/10.1075/ata.xiv.03ang

ANGELELLI, C.; JACOBSON, H. E. **Testing and Assessment in Translation and Interpreting Studies**. ATA Scholarly Monograph Series XIV. Amsterdam/Philadelphia: John Benjamins, 2009. https://doi.org/10.1075/ata.xiv

ARJONA-TSENG, E. A Psychometric Approach to the Selection of Translation and Interpreting Students in Taiwan. *In*: MOSER-MERCER, B.; LAMBERT, S. (Ed.). **Bridging the Gap:** Empirical Research in

Simultaneous Interpretation. Amsterdam/Philadelphia: John Benjamins, 1994. https://doi.org/10.1075/btl.3.08arj

BOGUCKI, L. (Ed.) **Teaching Translation and Interpreting:** Challenges and Practice. Newscastle upon Tyne: Cambridge Scholars, 2010.

DAVIES, A.; BROWN, A.; ELDER, C.; HILL, K.; LUMLEY, T.; MCNAMARA, T. **Dictionary of Language Testing**. Reprint. Cambridge: Cambridge University Press, 1999/2002.

DODDS, J. M. On the Aptitude of Aptitude Testing. **The Interpreters' Newsletter**, v. 3, p. 17-22, 1990.

DONOVAN, C. Experience with Interim Testing at ESIT – Some General Considerations, 2008. Available at https://www.emcinterpreting.org/resources/projects. Accessed in October 2019.

GAMBIER, Y.; GILE, D.; TAYLOR, C. **Conference Interpreting:** Current Trends in Research. Amsterdam/Philadelphia: John Benjamins, 1997. https://doi.org/10.1075/btl.23

GILE, D. L'évaluation de la qualité de l'interprétation en cours de formation. **Meta**, v. 46, n. 2, p. 379-393, 2001. https://doi.org/10.7202/002890ar

GILE, D. L'évaluation de la Qualité de L'interprétation par les délegues: Une Étude de Cas. **The Interpreters' Newsletter**, v. 3, p. 66-71, 1990.

HEWETSON, Z. Mid-year Test and Resits at the University of Westminster: an Update, 2008. Available at https://www.emcinterpreting.org/resources/projects. Accessed in October 2019.

HLAVAC, J. Formalizing Community Interpreting Standards: A Cross-National Comparison of Testing Systems, Certification Conventions and Recent ISO Guidelines. **International Journal of Interpreter Education**, v. 7, n. 2, p. 21-38, 2015.

JACOBSON, H. Moving beyond Words in Assessing Mediated Interaction – Measuring Interactional Competence in Healthcare Settings. *In*: ANGELELLI, C.; JACOBSON, H. E. (Ed.). **Testing and Assessment in Translation and Interpreting Studies**. ATA Scholarly Monograph Series XIV. Amsterdam/Philadelphia: John Benjamins, 2009. https://doi.org/10.1075/ata.xiv.04jac

KEISER, W. Selection and Training of Conference Interpreters. *In*: GERVER, D.; SINAIKO, H. W. (Ed.). **Language, Interpretation and Communication, Proceedings of the NATO Symposium on Language, Interpretation and Communication**. New York: Plenum Press, 1978. https://doi.org/10.1007/978-1-4615-9077-4_3

KURZ, I. Conference Interpreting: User Expectations. **ATA – Proceedings of the 30th Annual Conference**. Medford, New Jersey: Learned Information, 1989. p. 143-148.

KURZ, I. Conference Interpreting: Quality in the Ears of the User. **Meta**, v. 46, n. 2, p. 394-409, 2001. https://doi.org/10.7202/003364ar

LONGLEY, P. An Integrated Programme for Training Interpreters. *In*: GERVER, D.; SINAIKO, H. W. (Ed.). **Language, Interpretation and Communication, Proceedings of the NATO Symposium on Language, Interpretation and Communication**. New York: Plenum Press, 1978. https://doi.org/10.1007/978-1-4615-9077-4_6

MACDONALD, P. It Don't Mean a Thing... Simultaneous Interpretation Quality and User Satisfaction. **The Interpreters Newsletter**, v. 18, p. 35-59, 2013.

MACKINTOSH, J. A Review of Conference Interpretation: Practice and Training. **Target** , v. 7, n. 1, p. 119-133, 1995. https://doi.org/10.1075/target.7.1.10mac

MOSER, P. Survey on Expectations of Users of Conference Interpretation, Final Report in pdf available at https://aiic.net/page/736/survey-on-expectations-of-users-of-conference-interpretation/lang/1, 1995 (pdf version 2012). Accessed in November 2019.

MOSER-MERCER, B. Aptitude Testing for Conference Interpreting: Why, When and How. *In*: MOSER-MERCER, B.; LAMBERT, S. (Ed.). **Bridging the Gap: Empirical Research in Simultaneous Interpretation**. Amsterdam/Philadelphia: John Benjamins, 1994. https://doi.org/10.1075/btl.3.07mos

MOSER-MERCER, B.; LAMBERT, S. **Bridging the Gap: Empirical Research in Simultaneous Interpretation.** Amsterdam/Philadelphia: John Benjamins, 1994.

PELLATT, V.; GRIFFITHS, K.; WU, S.-C. **Teaching and Testing Interpreting and Translating**. Bern (Switzerland): Peter Lang, 2010. https://doi.org/10.3726/978-3-0353-0267-7

PÖCHHACKER, F. Quality Assessment in Conference and Community Interpreting. **Meta**, v. 46, n. 2, p. 410-425, 2001. https://doi.org/10.7202/003847ar

SAWYER, D. B. **Fundamental Aspects of Interpreter Education**. Amsterdam/Philadelphia: John Benjamins, 2004. https://doi.org/10.1075/btl.47

SELESKOVITCH, D.; LEDERER, M**. Pédagogie Rasionnée de L'interprétation**. 2eme edition, corrigée et augmentée. Paris and Brussels: Didier Érudition/Office des Publications Officielles des Commuautés Européennes, 2002.

SETTON, R.; DAWRANT, A. **Conference Interpreting: A Trainer's Guide**. Amsterdam/Philadelphia: John Benjamins, 2016. https://doi.org/10.1075/btl.120

SHLESINGER, M. Quality in Simultaneous Interpreting. *In:* GAMBIER, Y.; GILE, D.; TAYLOR, C. (Ed.). **Conference Interpreting:** Current Trends in Research. Amsterdam/Philadelphia: John Benjamins, 1997. https://doi.org/10.1075/btl.23.08shl

SIKORSKI, J. Interpreter Aptituded in Testing – Procedures. *In*: BOGUCKI, L. (Ed.) **Teaching Translation and Interpreting: Challenges and Practice**. Newscastle upon Tyne: Cambridge Scholars, 2010.

STEVENSON, D. K. Beyond Faith and Face Validity: The Multitrait-Multimode Matrix and the Convergent and Discrimant Validity of Oral Proficiency Tests. In: PALMER, A. S. et al. (Ed.) **The Construct Validation of Tests of Communicative Competence**. Washington, D.C.: TESOL, 1981.

VIDAJOVIC, I. and GALACZI, E. The Measurement of Speaking Ability 1913-2012. *In*: WEIR, C. J.; Peter J. GROOT, M.; TROSPER, G. A. **Measured Constructs:** A History of Cambridge English language examinations 193-2012. Cambridge: Cambridge University Press, 2012.

WEIR, C. J. An overview of the influences on English language testing in the United Kingdom 1913-2012. *In*: WEIR, C. J. et al. **Measured Constructs: A History of Cambridge English language examinations 193-2012.** Cambridge: Cambridge University Press, 2012.

WEIR, C. J. et al. Measured Constructs: **A History of Cambridge English Language Examinations 193-2012.** Cambridge: Cambridge University Press, 2012.

WU, S. C. Some Reliability Issues of Simultaneous Interpreting Assessment within the Education Career. *In*: PELLATT, V.; GRIFFITHS, K.; WU, S.-C. **Teaching and Testing Interpreting and Translating**. Bern (Switzerland): Peter Lang, 2010.

Recebido em: 02.12.2019 Aprovado em: 25.12.2019