Filozofická fakulta Univerzity Palackého

Criteria of Quality Assessment in Interpreting

(Bakalářská práce)

2016 **Karolína Tomanová**

Filozofická fakulta Univerzity Palackého

Katedra anglistiky a amerikanistiky

Criteria of Quality Assessment in Interpreting

Kritéria hodnocení kvality tlumočení

(bakalářská práce)

Autor: Karolína Tomanová

Studijní obor: Angličtina se zaměřením na komunitní tlumočení a překlad

Vedoucí práce: PhDr. Veronika Prágerová, Ph.D

Olomouc 2016

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a uvedla úplný seznam citované a použité literatury.

V Olomouci dne 3.5.2016

Chtěla bych poděkovat PhDr. Veronice Prágerové, Ph.D za ochotu, odborné vedení a cenné rady, které mi v průběhu psaní práce poskytla.

List of Abbreviations

AIIC – International Association of Conference Interpreters (Association internationale des interprètes de conférence)

CACL – a committee for admission and language classification of applicants under the AIIC

Table of Contents

Introduction

The issue of quality has been on interpreters' minds quite a bit in the last few decades and even though it is a very popular subject we still do not have an established definition of what constitutes "good" interpreting in a given setting. One of the branches of studies of interpreting quality is quality assessment. As Moser-Mercer puts it: "When speaking of quality we immediately think of how we can measure it, or evaluate, or assess it." (1996, 45) After conducting a review of literature on interpreting quality and interpreting quality assessment, defining assessment and considering different assessing mechanisms, I will attempt to answer the question of who should execute such an assessment. Is it a fellow interpreter? Is it the "user", the person listening to an interpretation? Or someone else completely? Those are the questions I will be asking in the first part of the thesis. The aim of the middle part will be to try to come up with a sort of a "recipe" for interpreting quality researchers on how to make a questionnaire to assess a given interpreting performance. I would like to offer a few practical tips and guidelines on designing a quality assessment questionnaire. In the last part, a set of both linguistic and non-linguistic criteria will be listed with an emphasis on who should assess or rate a given interpreting performance according to these criteria. Criteria will be organized in a table where on one side, there will be criteria better suited to be assessed by the user and on the other criteria intended for fellow interpreters. While I try to focus on both simultaneous and consecutive interpreting, literature concerned with interpreting quality is available mainly for simultaneous conference interpreting. I chose this topic because I find something missing from most studies on quality assessment that have been published so far. It can be missing information about the respondents or the particular questionnaire design. Authors often describe what they asked the respondent but rarely do they reveal how exactly they did it. And that is why it can sometimes be difficult for researchers to simulate the conditions of the original survey which, in my opinion, disqualifies the assessment grid to be reusable which I find to be one of the most important characteristics of an interpreting quality assessment tool.

1. Approaches to Assessing Quality

1.1 Why quality

As Chiaro and Nocella point out in their paper: “Systems of quality control are gradually becoming common practice in every imaginable walk of life” and interpreting can no longer be an exemption. (2004, 278) The problem is we have not yet agreed on a generally applicable definition of what quality is. (Kalina 2005, 768) So how can we measure something we do not know what is? Because of the demands of today’s market we should try. There are many reasons to measure quality and also many parties potentially interested in doing so. Whether it is the interpreters themselves wondering about their own performance, intermediaries having questions about the interpreters they hired, interpreters’ booth mates wanting to know how the other interpreter dealt with a particular situation, or experimenters manipulating certain variables and observing quality as a dependent variable. (Moser-Mercer 1996, 46)

1.2 Assessment/measurement/evaluation

Moser-Mercer goes on to differentiate between evaluating, measuring and assessing quality. She suggests the term “evaluate” to apply to quality in the field, “measure” for laboratory experiments where a specific type of scales is used and “assess” in the case of observing the change in quality over a given period of time. (1996, 47) After reviewing literature on the topic of quality in interpreting, it was decided to use the term “assess” for all three as this term prevails.

1.3 Expectations vs. actual assessment of the end product

One of the first to point out the need for distinction between these two was Franz Pöchhacker (1994, 233). Pöchhacker and Shlesinger define these two activities as follows: “While research on hypothetical expectations yields data on what users think they want, quality assessment studies are needed to demonstrate what really shapes users’ appreciation of an interpreter’s performance.” (2002, 297) In her study on importance of nonverbal communication, Collados Aís concludes that these two do not necessarily correspond. It turns out that while the users might theoretically consider certain aspects of interpreting very important, they later focus on different aspects when they are asked to rate an actual performance. (1998/2002, 328-336)

Unfortunately, several studies in quality assessment do not highlight which approach they have chosen. Some of the founding studies to have chosen to explore users' expectations were Hildegund Bühler's study from 1986 and Ingrid Kurz' from 1993. Examples of assessment of the end-product are Henri Barik's error count from 1971 and recently also research of Elisabet Tiselius.

1.4 Componential, non-componential or an error count?

Tiselius differentiates between componential and non-componential approach to quality assessment. Componential research evaluates interpreting from the point of view of its different components. "Components typically cover fluency, correctness (terminology, grammar, syntax), sense consistency (with original), logical cohesion, intonation, accent, style and more (or less)" (Interpretings.net 2013) A large portion of quality assessment research has been componential so far. Is it because we consider it more intuitive? Both Elisabet Tiselius and Collados Aís point out that the components can affect each other where, for example, foreign accent can result in a lower score for fidelity. (Interpretings.net 2013)

Keeping in mind the purpose of quality assessment, which is being able to tell whether a given interpreting event was successful or unsuccessful, the most efficient way to assess would be to look at one factor, or component, and be able to assess immediately. That is where non-componential way of assessment comes into play. Of course, no assessment device is perfectly non-componential so far, it is just a tendency. We can say that a method is of non-componential nature but no method is entirely non-componential just yet. Even the part of research by Tiselius she calls non-componential consists of two components.

A line of research called error counting could be considered a potential third approach to quality assessment. As much as the idea of non-componential assessment sounds efficient, there is probably a good reason why we have been opting for componential methods so frequently instead.

1.4.1 Componential research

1.4.1.1 Hildegund Bühler

Along with Ingrid Kurz, Hildegund Bühler's research is considered ground work in quality assessment. Both surveys were componential and were based on asking either fellow interpreters or delegates what attributes of interpreting they consider to be significant. Bühler was researching what criteria AIIC members consider important when assessing performance of a potential AIIC member. She introduces 16 criteria—eight linguistic, seven extra-linguistic and one concerning positive feedback from the delegates. (Bühler 1986, 231)

1.4.1.2 Ingrid Kurz

Kurz took eight of the criteria Bühler introduced and applied them to conference participants instead of fellow interpreters. Her research is groundbreaking because she was the first one to ask the target audience instead of just asking colleagues. Her findings suggest that the importance of different criteria indeed varies when it comes to conference participants and fellow interpreters. She comes to the conclusion that different user groups bear different expectations of interpreting.

1.4.1.3 Delia Chiaro and Giuseppe Nocella

Italian researchers Chiaro and Nocella also studied the importance interpreters attribute to different aspects of quality. Unlike Bühler and Kurz they administered the questionnaire online and their research was significant because they studied a large sample of 286 interpreters.

1.4.1.4 Ángela Collados Aís

Research of Collados Aís is very significant because she compared the findings of an expectations survey with those of an actual assessment of a performance. While this has been done before, she was the first one to do both on one set of recorded performances. They were videos with simultaneous interpretation as a voice-over. She first asked 42 legal experts to fill in an expectations questionnaire, then she asked them to watch three video recordings of interpreting and then they filled in another questionnaire evaluating the performance. (1998/2002, 330) Macarena Pradas Macías followed this research design in her study concerning fluency in interpreting. (2006, 28)

1.4.1.5 Elisabet Tiselius

Tiselius uses and adapts an assessment model created by Adelina Ivanova. This approach was first used to observe the progress in interpreting quality between novices and expert interpreters. Ivanova's research focuses on the processes taking place in the course of an interpreting event but Tiselius takes her method and turns it into a quality assessment device. Tiselius divides Ivanova's criteria into three categories: processing problems, monitoring observations and strategies. A transcription of an interpreting performance gets coded and tagged according to Ivanova's criteria. (Tiselius 2005, 15)

1.4.2 Non-componential research

1.4.2.1 Elisabet Tiselius

Another approach adapted by Elisabet Tiselius are Carroll's Scales. In her assessment of transcriptions of interpreting Tiselius uses scales originally developed by John B. Carroll for assessing quality of machine translation. She adapts the scales to be used for interpreting. On one side, there is "intelligibility" which indicates if "the translated text can be read and understood independently of other texts or sources of information." Ideally, we should not be able to tell that a text is not an original and that it is a translated text. On the other side, there is "informativeness" which shows whether there is any difference in the amount of information provided by the original and information provided by the translated text.

> *"...an interpretation faithful to the original and easy to understand, a good interpretation, should score high for intelligibility and low for informativeness, since informativeness is the measure for information added to the interpretation, not present in the original."* (Tiselius 2005, 32)

The usability of Caroll's original scales was criticized by Stenzl. She said that such methods focus too much on the target text and not as much on the transfer and while lack of intelligibility might really indicate there is a problem in the transfer, it does not work the other way around. Intelligibility is no proof of fidelity to the source text. (Stenzl 1983, 22)

1.4.3 Error count

1.4.3.1 Henri C. Barik

Barik works with transcripts of translations similarly to Elisabet Tiselius. The transcripts are tagged by "judges" and the data is then processes in terms of a number of instances per a 100 words. This number refers to instances of omission, addition, substitution and error. The amount of material that has been omitted, added, substituted, or was transferred falsely, on the other hand, is expressed in percentage. Barik warns that:

> *"(the system is not) intended to reflect except in a very gross way on the adequacy or 'quality' of an interpretation since other critical factors such as delivery characteristics: voice intonation, appropriateness of pausing, etc., are not taken into consideration."* (Barik 1971, 207)

Moser-Mercer believes that quality in interpreting cannot be measured without using error scales. (1996, 51) The usability of this method has since been questioned by Gerver and Stenzl as mentioned by Gile (1994, 47)

2. Who Should Assess

There are several parties involved in the process of interpreting and many of them can form an opinion as to what quality in interpreting means. As Garzone puts it:

> *"The basic problem is that quality is the sum of several different, heterogeneous aspects, some of which involve different subject —interpreters, clients, users, speakers —each with a different view and perception of quality..."* (2002, 107)

Researchers generally only use either the target audience (or "users") who the interpretation is intended for or fellow interpreters as respondents. There are positives and negatives to both points of view. So far, the opinions as to who should assess interpreting have been very polarized. Researchers either chose their colleagues as their respondents and had very strong arguments backing up their decision or they chose the users. Cartellieri observes that "very often, a good interpreter is two quite different people, being one thing to a conference participant and another to a colleague." (1983, 213) Either way determining the assessors and identifying them in terms of what they are basing their judgements on should be the building block of any credible research into interpreting quality. (Kellett Bidoli 1996, 122)

2.1 Users of interpreting services as assessors; pros and cons

The very reason for existence of interpreting is the person in need of help when it comes to interlingual communication and it only makes sense that we would adapt the quality requirements to the user of interpreting services. Sylvia Kalina even says that: "listeners' needs and expectations are central to the discussion of quality". (2005, 774) Kotler and Armstrong state this marketing principle as follows: "Quality must begin with customer needs and end with customer perception." (Kotler and Armstrong 1994, 568) In her comment following Hildegund Bühler's pioneering study, Danica Seleskovitch notes on the topic of fluency in an interpreting performance that "interpretation should always be judged from the perspective of the listener and never as an end in itself." (1986, 236) Ingrid Kurz comes to the conclusion that: "They [her findings] clearly show that the target-language receiver or listener must be seen as an essential element in the process.” in her study of different user groups. (1993, 20)

On the other hand, the user cannot understand the source language in most cases so they can easily be misled by smooth delivery of the interpretation. That means that the user can be under a false impression that the interpretation was of a “high quality” or “correct” when, in fact, the message could be distorted or parts of the interpretation even missing. (Shlesinger 1997, 127) Sylvia Kalina was interested in the same problem and she noted that the user assesses the interpretation according to standards of monolingual communication. (2005, 775) Bühler warns that delegates are “likely to judge the quality of interpretation by such superficial criteria as native accent, pleasant voice, and fluency of delivery. (1986, 233) It is the criterion of “fidelity” (or sense consistency) that they cannot judge. Unless the interpreter makes a significant error the user has no way of noticing an inconsistency with the source text. (Gile 1991, 196) Ángela Collados Aís noticed that when asking the end-user to evaluate according to the criteria which required knowledge of the source language, they seemed unaware of this limitation and evaluated the interpretation according to these criteria anyway. Only a few subjects noted down that they are not fit to evaluate this. (2002, 335) She even went on to say that: “The users are not good judges of quality, simply because they are not in a position to perform this task.” (2002, 336) Chiaro and Nocella call the end-users “people who consume the service but know very little about it” and come to the conclusion that the user can judge the quality of the interpretation only partially. (2004, 281) Shlesinger asks a simple question: “Do our clients know what’s good for them?” (1997, 126)

3. The Process of Creating a Questionnaire

3.1 Before we start

There are many important decisions to make before we even begin picking which criteria we are going to use in a given survey. First of all, we have to decide whether to focus on the interpretation as a communicative interaction and observe the entire process from the client, the speaker, the interpreter, the listener to even the fellow interpreters. Or, we can have a more product-oriented approach and only focus on the portion of the process that we can record, that means the interpretation itself. This classification was introduced by Pöchhacker. (2001, 412) For the purposes of this thesis, it was decided to focus on the product itself. Pöchhacker points out that

> *"Whereas such effect-oriented studies can do without the analysis of textual-linguistic data, performance-oriented experimentation has traditionally been associated with the processing of recordings and transcripts."* (2001, 419)

At this point of conducting research, we have to decide whether we want to study expectations the different participants have of interpreting or whether we want to assess an actual performance. This decision affects the choice of criteria and also the final form of the items in the study. Criteria for both performance assessment and an expectations study will be listed. Once we have made that decision, it is time to choose what the study will look like. If we opt for an expectations study, the most common research design is the componential one, usually in a form of a questionnaire with either open or closed questions. Hale and Napier remind that answers from closed questions can be easily quantified at the time of the analysis. They recommend for open questions to serve as tools to get commentary on top of the answers to the closed questions. Additionally, they warn us that if a majority of items in our questionnaire are open questions then a questionnaire might not be the correct tool for the given research. (2013, 57) Instead of a questionnaire, interviews can also be conducted as long as the number of respondents is not too large. In the case of a performance assessment, there are more options. We can conduct an error-count following in Barik's steps, non-componential research as originally done by John B. Carroll, or we can administer a questionnaire consisting of different items i.e. componential research. A suggestion of the mental processes taking place before designing a survey was outlined in ***Figure 1***.

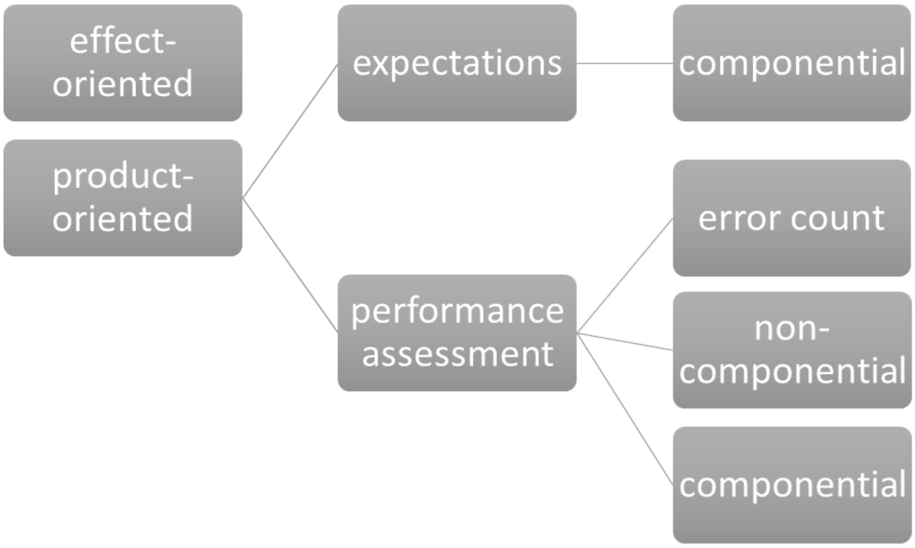

Figure 1. *Decisions taking place before designing a survey.*

At this point of the survey, we have to make sure that a questionnaire is the way to go even though it might be the most popular research tool in interpreting quality. Hale and Napier define a survey as “a means of gathering information from particular sources to ascertain certain situations or conditions” but they admit that surveys are most commonly associated with questions. The different kinds of surveys are listed: observations, text analyses, interviews, focus groups and questionnaires. (Hale and Napier 2013, 52) Criteria listed in this thesis are primarily intended for use in questionnaires. Hale and Napier recommend that interviews and focus groups be used only if we are looking for detailed answers. (2013, 52) Interviews and focus groups can be very useful tools as a part of the pre-processing stage, the initial stage before administering a questionnaire where we can make sure what items to include in it and making sure all of our questions are appropriate for the chosen sample. Chiaro and Nocella claim to have held “endless brainstorming sessions” and interviews with professional interpreters before they decided which criteria were involved. (2004, 283) Aside from use in the initial stages of research (Hale and Napier highlight the usefulness of interviews and focus groups at the beginning of research especially when we are studying a topic that is new to us), they also suggest using it after administering a questionnaire to ask respondents to clarify some of the responses provided. (2013, 53) It might save us from not being able to analyze some of them. Chin Ng held interviews with respondents a week after the questionnaire stage of her research on speech levels in Japanese. First, she gave the subjects a briefing on the linguistic terms used, then she played them parts of the recorded interpretation performances which the respondents were asked to listen to in the questionnaire stage just to jog their memory. After that they were asked a few questions which focused on more specific issues then the questionnaire and asked to rate the interpreters' overall use of speech levels. (Chin Ng 1992, 37)

3.2 Sampling aka Who Should Assess What

It seems that the question we should be asking is not "Who should evaluate" but it is "Who should evaluate what?". Why ask those end users who do not speak the source language to form an opinion on, for example, *sense consistency with the original* when it is clearly only someone who speaks the source language that can evaluate this criterion. Why ask the same users about the rate of success of the transfer altogether when they will just easily be swayed by the properties of the delivery. But as it was pointed out before, we cannot forget that the end user is the very reason for interpreting so they need to play one of the key roles when it comes to assessing interpreting quality. On the other hand, there are interpreting features such as *accent, pleasant voice, intonation,* or *diction* that colleagues from the interpreting profession can assess just as easily as the end users. But if we were to opt for a research design where we have to either choose the interpreters or the users as assessors for a given criterion, let us bear in mind that while the interpreters' opinion is just as legitimate as the users', it is the end user that the interpretation is intended for.

The eternal issue of who should assess an interpreting performance is associated with the issue of sampling in a survey, it seems. Dörnyei defines the sample as follows

> *"The sample is the group of participants whom the researcher actually examines in an empirical investigation and the population is the group of people whom the study is about."* (2007, 96)

Sampling methods are divided into probabilistic and non-probabilistic. The difference lies in the probability of being chosen as a respondent from a given population. While in a probabilistic sample the probability is higher than zero, with a non-probabilistic method the sample is up to the researcher. That means that we are unable to determine the probability of being chosen or it is zero. Hale and Napier associate quantitative research with probability methods and qualitative with non-probability. (2007, 68-73) Yet it seems that most research in interpreting quality uses non-probabilistic samples as a lot of times questionnaires are the research tools of choice. As listed by Hale and Napier, probability sampling techniques include:

- Simple random sampling – comprises of the lottery method, the systematic method where we, for example, put together a list of a 100 and pick every tenth person on the list to be a respondent (in this case the sample would be 10% of the population), and the random number method where a special computer software picks a random sample (this method is good especially for a large number of respondents);

- Stratified sampling – when we divide the population into subgroups to make sure each of these groups gets represented (for example, by language or country) and then use one of the techniques of simple random sampling;

- Cluster sampling – the population gets divided into subgroups again but this time, for example, by the region (such as Northern, Southern, Western, and Eastern Europe) and then one of the probability methods is applied.

These are the techniques of non-probability sampling:

- Quota sampling – similar to stratified sampling but without the randomness, the researcher picks the participants from the subgroup;

- Network or snowball sampling – members of a professional organization are contacted and asked to pass the survey onto their colleagues;

- Convenience or opportunity sampling – the researcher uses his colleagues, fellow students, family, or friends to be respondents. (Hale and Napier 2013, 68-73)

It seems that interpreting researchers have been favoring the last two techniques. Bühler addressed her questionnaire to the members of the AIIC and also the CACL committee of the AIIC and distributed at a conference in Brussels assuming that this sample reflects needs and expectations of the users as well. (1986, 231-234) This assumption was later contested by Kurz. (1993, 14) Chiaro and Nocella sent their online questionnaire to "interpreters belonging to several professional associations". (2004, 284) Kurz went in a different direction and administered hers at the occasion of several conferences on different topics like medicine, quality control, and a Council of Europe meeting. (1993, 15) Collados Aís et al. relied on their colleagues from other universities as respondents for their study on simultaneous interpreting. (2007, 7) Both Chin Ng (1992) and Pradas Macías (2006) talk about the importance of choosing respondents with previous experience with conference interpreting. Pradas Macías notices a trend already pointed out by other researchers where first time users of interpretation tent to rate the performance less favorably than experienced users. (2006, 38) Researchers of interpreting quality do prefer to use their own colleagues, fellow students and families as respondents but one cannot blame them considering how relatively small the interpreting community is especially if different surveys require different knowledge of languages of the respondents. Hale and Napier comfort us by saying that even though a sample gained through a non-probabilistic method cannot be considered a representative sample, the results of the survey can be extended to other studies provided that it is well devised and well thought out. (2013, 71)

3.3 Questionnaire design

3.3.1 Participant information page

Once we know exactly what kind of research we wish to conduct, we can get underway and design the physical questionnaire. One section every questionnaire should have is the participant information page. The need for this page arises from the fact that any participation from the side of the respondent has to be voluntary. It should contain all the information the respondent needs to make the decision whether they want to participate. At the end of the page, there can be a line which states that by completing the questionnaire the participant gives an informed consent. Different universities may have different requirements of what to include on this page. (Hale and Napier 2013, 55) In their survey on obtaining feedback from non-English speakers, Garber and Mauffette-Leenders included basic information about the interpreting agency and what was important also instructions as to how to mail the questionnaire back to the agency on the participant information page. (1997, 141)

3.3.2 Types of questions

Hale and Napier divide questions into three categories. The first part of the questionnaire is formed by factual questions. These can, for example, concern participant's age or qualifications. (2013, 58-59) Information gained through this part of the questionnaire can later serve when we describe the sample just as Chiaro and Nocella did in their study. (Chiaro and Nocella 2004)

The second group of questions are behavioral questions. When administering a questionnaire to interpreters, Hale and Napier list asking how many times the interpreter worked in different settings as an example. (2013, 60) Garber and Mauffette-Leenders ask "How many times have you had this interpreter before?" (1997, 141)

Attitudinal questions are the third and last category. These are designed to find out the respondents' opinions. They range from the least leading alternatives like open questions to the most leading ones such as lists of statements with ranges of how much the respondent agrees. (Hale and Napier 2013, 60-61)

In the questionnaire administered by Collados Aís et al. the first two types of questions are missing. (2007, 247-255) The authors could have been trying to keep the questionnaire short so they went straight to the questions essential to the research. Hale and Napier recommend that the maximum amount of time needed to complete a survey should not exceed 30 minutes with 10-15 minutes being ideal. They point out that after 15 minutes some people might even decide to abandon the questionnaire which would have an impact on the results. (2013, 55)

3.3.3 What to avoid

Hale and Napier remind us the things to avoid when writing the questions. We have to ask about concepts which are familiar to the participant while using terms they know. This is the same issue Collados Aís was dealing with in her questionnaire when she asked users of interpreting services to assess interpreting according to criteria such as *sense consistency with the original* and *completeness of interpretation* which was already mentioned. The questions were asking something the respondents had no chance of knowing since they were unfamiliar with the source language and thus had no way of checking. (2002, 335) Another thing to avoid are double-barreled questions, questions asking or containing two concepts. Questions also should not be vague and ambiguous, we should try to avoid using passive voice and they should not contain embedded clauses. There are two types of questions we should avoid and those are loaded and leading questions. But most importantly, we need to be asking questions that are relevant to our research. (Hale and Napier 2013, 63)

3.4 Administration

As to administering the questionnaire, Hale and Napier recommend doing so online although they warn that some respondents may not be comfortable enough using technology and might refrain from taking the survey which would skew the sample. (2013, 74) Out of the seven major surveys listed in this thesis, only two were administered a questionnaire online. Chiaro and Nocella's survey was entirely web-based created in a software called Microsoft Personal Web Server. (Chiaro and Nocella 2004, 284) AIIC used a questionnaire generator tool LimeSurvey for their web-based research. (2008) The relatively low number of web-based surveys is either because of how old the surveys are or, as in the case of Kurz, paper was probably also more convenient since she administered it at conferences. (Kurz 1994) Chin Ng went another way and trusted the respondents to complete the survey since it was an experiment and they already agreed to participate. She gave them a list of instructions, asked them to listen to recorded interpretations and answer her questions. (Chin Ng 1992, 37) Collados Aís (1998) and Paradas Macías (2006) used video recordings in their research and they both highlight the importance of microphone distance and voice-over level in the experimental videos. (Pradas Macías 2006, 30) The AIIC highlights the importance of anonymity of responses in their web-based survey, they created two separate databases for this purpose, one containing all of the e-mail addresses and the other containing the responses. The two databases were not linked in any way so that the administrator had no way of seeing the response of the respondent. (AIIC 2008)

3.5 Piloting

Piloting a questionnaire means testing it before it is administered in its final form. Hale and Napier urge us to check "typographical errors, clarity of questions, missing questions and how long it took them (respondents) to complete it." (2013, 67) Chiaro and Nocella piloted their questionnaire in its web-based form already for a few days before launching the final version. (2004, 284) The AIIC contacted 36 professional interpreters not affiliated with AIIC as the questionnaire was intended for AIIC members. (2008) Pradas Macías conducted three preliminary tests as her research was based on video recordings as "stimulus material" so this technical aspect of her work was very important. However, her preliminary tests were not conducted as a form of piloting the study but they took place in order to design the experiment in the first place. She needed to know where interpreters would make silent pauses when interpreting given material. The only one of her three preliminary tests that could be considered a piloting run would be the third test where she made sure that the results of her experiment would not be distorted by anything else besides the effect of the silent pauses which were the subject of the study. The respondents were asked to watch an interpretation containing a high number of manipulated silent pauses and fill in a listening comprehension test. (Pradas Macías 2006, 30)

3.6 Return rate

When talking about return rate, one needs to distinguish between the different ways the questionnaires get distributed. The best return rate is in the case of experiments. Once the participant agrees to partake, they usually complete the questionnaire. The return rate for experiments would be 100% or close to 100%. Examples of experimental studies are Chin Ng (1992), Collados Aís (1998) and Pradas Macías (2006). Then there is the case of administering questionnaires at conferences where the authors usually do not state the rate but only the number of questionnaires handed in such as Bühler (1986), and Kurz (1993). Then there are questionnaires that get mailed to a large number of potential respondents with usually only a smaller number of people deciding to partake in the survey. This is when it makes sense to indicate the return rate. Chiaro and Nocella e-mailed approximately 1000 invitations with a link to the survey, out of which around 200 were returned to the sender because of delivery errors. The total sample number was 286 and the authors stated that the return rate doubled that of traditional surveys which is usually 10-15% as they state. Since 286 is 35.75% of 800, it is unclear if maybe they did not include all the data needed. (Chiaro and Nocella 2004, 284) Čeňková's survey intended for interpreters from the Czech Republic had a return rate of 15%. (1998, 163) It seems that 15% is a common return rate for questionnaires sent to a larger number of potential respondents.

3.7 Analysis

3.7.1 Natural sciences approach vs. social sciences

When analyzing an interpreting survey, we can either go the way of social sciences or follow in the steps of natural sciences. The biggest promoter of the exact natural sciences approach is Franz Pöchhacker followed by Delia Chiaro and Giuseppe Nocella. Ingrid Kurz and Sylvia Kalina can be considered more on the social sciences side.

3.7.2 Quantitative or qualitative research

The difference between quantitative and qualitative research is something that we have to keep in mind already when designing a questionnaire but it plays a big role when we analyze the results of our surveys. Hale and Napier note that "under quantitative methods, concepts are converted into numbers and are quantified." With qualitative methods, on the other hand, "the quality or characteristics of concepts are explored, described and interpreted." (Hale and Napier 2013, 15) In quantitative research, statistics is used while qualitative research uses personal descriptions of phenomena and speculates on conclusions. Hale and Napier highlight that most research will have both quantitative and qualitative characteristics. (2013, 16) If we are conducting qualitative research with a small number of respondents and simple questions, we can analyze the results manually. To analyze research with a large sample and complex questions, however, we will need a software package used for data analysis such as SPSS or nVivo. (Hale and Napier 2013, 75)

3.7.3 Non-parametric and parametric data

Data generated by questions in our questionnaire will either be non-parametric or parametric. Non-parametric (nominal or ordinal) data refers to names and categories which can be coded into numbers to make analysis easier. Hale and Napier listed semantic scales as an example. We can attribute (1 – 2 – 3 – 4 – 5) to ("completely useless" – "useless" – "neutral" – "useful" – "very useful") respectively if we were to rate the effectiveness of an interpreting teaching method, for example. But we have to remember that the differences between rankings are not equal so it cannot be said that "completely useless" is half as useful as "very useful", they are not exact measurements. These are not quantitative data and we cannot treat them as such. (Hale and Napier 2013, 76) And this is what Chiaro/Nocella (2004, 283) and Pöchhacker (2005, 147) do not like about the surveys of Bühler (1986) and Kurz (1993). Pöchhacker sees the same problem Hale and Napier described in that Kurz used arithmetical mean to analyze her ordinal data. He felt that she should have described the results using percentages. Here, too, it is a matter of distance between the ratings. On the scale of ("highly important" – "important" – "less important" – "irrelevant"), the intervals between the items are not the same. (Pöchhacker 2005, 147) Parametric data (ratio or interval) refers solely to numerical answers with the numbers being exact. (Hale and Napier 2013, 76)

3.7.4 Descriptive vs. inferential statistics

If we are only looking to do a descriptive, qualitative study then tools such as frequencies, percentages, and means (averages) will do. This is called descriptive statistics. But if we are looking to conduct a quantitative study and want to see if the results are statistically significant in any way, we will have to use inferential statistics. For these purposes we will have to use tests with the help of the above mentioned data analysis software packages. According to what type of data we are dealing with, we choose the appropriate test:

- T-test – for quantitative data to see whether two sets of scores are significantly different;
- ANOVA (analysis of variance) – for quantitative data to see whether more than two sets of scores are significantly different;
- Pearson's correlation – for quantitative data to see if there is a correlation between two variables;
- Chi-squared – for nominal data to see if the distribution of frequencies is not accidental. (Hale and Napier 2013, 78)

Bühler (1986), Collados Aís (1998), and AIIC (2008) simply described their results with the help of percentages. Kurz (1993) used the mean as already pointed out by Pöchhacker. He later took Kurz' data and performed an analysis himself doing crosstabulation with a chi-square test which he decided was not appropriate because of the low number of respondents. He also conducted the Kruskal-Wallis test which he described as a "nonparametric equivalent of the one-way analysis of variance (ANOVA)" and the Mann-Whitney test. (Pöchhacker 2005, 151-154) Pradas Macías (2006) used percentages for respondents' ratings of the influence of her quality parameters. In the second part of her study she used the mean but here it was justified—the respondents were asked to judge simultaneous interpretation they heard from a recording by assigning a value to each of her parameters using numbers from 1 to 5 with 1 being the most negative. (Pradas Macías 2006, 33) Chin Ng ranked interpreters according to their skills and competence in mastering speech levels in Japanese as expressed by her respondents. (1992, 38-39) Chiaro and Nocella used software called Statistica and performed multidimensional scaling which is a technique that helps the researcher develop certain "maps" or geometric representation of the results. (2004, 284)

4. Criteria

Criterion, or “an evaluative statement” as Pöchhacker (2004) calls it, is a standard according to which we can evaluate or assess interpreting. The rankings of these criteria were acquired simply by observing which ones had the highest percentage in importance in different studies. Criteria are differentiated from the rest of the text by cursive. At the end of the description of each criterion, a suggestion is offered as to who can assess according to it. May the list of criteria serve as an answer to the question of “Who should assess interpreting quality?”—the answer might just be that different agents in the process can assess according to different criteria and that the ideal assessment would be a fusion of ratings from fellow interpreters and end users. “Who should assess what?” might just be the question we should be asking ourselves.

4.1 Accent

The criterion of accent was used in the original studies of Bühler (1986, 232) and Kurz (1993, 15) and other studies followed: Chin Ng (1992, 38), Collados Aís (1998/2002, 331), Čeňková (1998, 164), Chiaro and Nocella (2004, 287), Pradas Macías (2006, 33) and the AIIC survey (2008). The criterion generally ranked low in importance or as “least” or “not important” in expectation surveys. Bühler suggested that the criterion bears a different significance depending on which language we are interpreting into and which from hinting at how French speakers are less tolerant and less used to hearing non-native accent in their language as opposed to, for example, English speakers who encounter it on regular basis. (1986, 232) She also speculated that *accent* is one of the superficial criteria delegates who are unable to understand the original are likely to judge the performance by. (1986, 233) Kurz’s study did not confirm this but admitted that in types of interpretation other than the ones she studied, such as media interpreting, it might be considered more important. (1993, 20) One of the respondents in the study of AIIC mentioned something called *native intonation* and pointed out that *accent* on its own is not as important but *native intonation* is essential to meaning. (2008) *Accent* can be assessed by both end-users and fellow interpreters, in both an expectation type of a survey and an actual performance assessment.

4.2 Pleasant voice

This criterion was used in Bühler (1986, 232), Kurz (1993, 15), Collados Aís (1998/2002, 331), Čeňková (1998, 164), Chiaro and Nocella (2004, 287), Pradas Macías (2006, 33), AIIC survey (2008) and ranked low or medium in importance right after *accent* or as "less important". It is an extra-linguistic criterion. Can also be assessed by both interpreters and end-users in both types of surveys.

4.3 Fluency of delivery

Fluency was used in the same studies as the previous criterion and it ranked medium to high in importance or as "highly important". In connection with this criterion, Bühler mentions the "charity principle" where the assessors of an interpreting performance are willing to overlook what the performance lacks in fluency as long as the sense of the original message is transferred, the interpretation makes sense (logical cohesion), and the interpreter uses correct terminology. (1990, 541) To the subject of *fluency*, she also notes that: "delegates may tire more easily and tend to feel uneasy if they have to listen to a halting interpretation, even if it is logically coherent and correct as to grammar and terminology." (1986, 232) Respondents in the survey of Čeňková call *fluency* a must. (1998, 167) Again, it can be assessed by both the end-users and the interpreters in both types of surveys. One can also talk about *speech rate* when talking about *fluency*.

4.4 Logical cohesion

The criterion of *logical cohesion* was used in the same studies as the three previous criteria. It ranked high in importance or as "very important". In her study, Bühler noticed a comment from one of her respondents that defined *logical cohesion* as IQ + education. She deduced that the criterion can be considered a prerequisite for communication in general, including monolingual. (1986, 232) It can be assessed by both interpreters and users either in expectation surveys or performance assessments.

4.5 Sense consistency with original message = FIDELITY

It is also known as *fidelity* and it is one of the original criteria interpreting researchers were interested in from the very beginning. *Sense consistency* ranked the highest in importance and as “most important” as rendering the sense of the original message is the very reason for existence of interpreting. It was used in the same studies as previous four. Bühler noted that if *logical cohesion* is a prerequisite for monolingual communication then *sense consistency* would be a prerequisite for interlingual communication. (1986, 232) Can be only assessed by interpreters and people speaking the source language in a performance assessment. In an expectation survey it can be ranked by both people who understand the source language and those who do not.

4.6 Completeness of interpretation

This criterion was again used in the same studies and it ranked medium or as “important”. It is one of the criteria from the survey of Collados Aís that the assessors rated even though they did not understand the source language. (1998/2002, 335) In Bühler’s study the more experienced in interpreting the respondent was, the lower the rating of importance of *completeness*. She theorizes that this is because interpreters find completeness of rendering over extended periods of time unrealistic. Or, it could be because these interpreters believe that the role of an editor in interpreting is an important one. (1986, 232) In performance assessments, the criterion can only be assessed by interpreters or speakers of the source language. Both speakers of the source language and those who do not speak it can rank it in expectation surveys.

4.7 Correct grammatical usage

The criterion of *grammar* is in one way or another represented in the studies of Bühler (1986, 232), Chin Ng (1992, 38), Kurz (1993, 15), Čeňková (1998, 164), Chiaro and Nocella (2004, 287) and AIIC (2008). It ranked medium to high in importance depending on the group of respondents. In her study, Kurz noticed that there were differences between her three user groups—the interpreters put the most importance on *grammar*, delegates were in the middle and the group of engineers the least. (1993, 19) Chin Ng found there to be a difference between how the criterion was rated by men and women. Women found it to be of high importance whereas men commented only on a few mistakes made in the interpretation. (1992, 38) Čeňková and also other researchers noticed that minor slips in grammar on the interpreter's side often go unnoticed by the audience. (1998, 167) *Grammatical usage* in an interpreting performance can be assessed by both end-users and interpreters. In expectations surveys anyone can have an opinion whether the criterion is important when it comes to interpreting.

4.8 Use of correct terminology

Correct terminology was a criterion in the studies of Bühler (1986, 232), Kurz (1993, 15), Collados Aís (1998/2002, 331), Čeňková (1998, 164), Chiaro and Nocella (2004, 287), Pradas Macías (2006, 33) and AIIC (2008). Chin Ng was interested in the *use of vocabulary*. (1992, 38) It ranked medium to high in importance or as "very important". Bühler mentions that *terminology* is connected to the criterion of *sense consistency* since the interpreter needs to use correct terms to render the meaning of the original message successfully. (1986, 232) Can be assessed by both fellow interpreters and end-users in a performance review. Both types of respondents can also rank its importance in an expectation survey.

4.9 Use of appropriate style

Used as a criterion by Bühler (1986, 232), Chin Ng (1992, 38), Collados Aís (1998/2002, 331), Chiaro and Nocella (2004, 287), Pradas Macías (2006, 33) and AIIC (2008). *Style* ranks in the middle for importance or as "important". Chin Ng's research is about the correct use of speech levels in interpretation into Japanese which is directly related to style as it deals with politeness. (1992, 35) The *use of appropriate style* can be assessed by both interpreters and end-users in both kinds of surveys.

4.10 Thorough preparation of conference documents

Bühler included this item in her original study but none of the researchers who adopted her criteria with the exception of Chiaro and Nocella (2004, 288) adopted this one. This is Bühler's second extra-linguistic criterion. It is a criterion which can only be used in expectation surveys. The term is better off used in the field of quality assurance than in assessment. Kellett Bidoli categorizes it as a “task-related factor”. (1996, 135) *Preparation of conference documents* ranked high in importance in Bühler’s survey. (1986, 232) Čeňková mentions *conference materials* as well in her study and uses it as a criterion but she means documents that are distributed to the users of the interpretation. Her survey concludes that the audience prefers to have the materials several days in advance before the event. (1998, 167)

4.11 Endurance

Another of the set of Bühler’s extra-linguistic criteria and it was ranked as “highly important” or “important”. (1986, 232) It is debatable whether it can be assessed only in expectation surveys or if it can be somehow measured in a performance assessment conducted by either fellow interpreters or end-users. Moser-Mercer tested *resilience* in her interpreting class. She considered sings of fatigue after 10 minutes of a class to be a handicap. (1985, 98) Chiaro and Nocella adopted the criterion of *endurance* in their expectations study and it ranked as the fourth most important extra-linguistic criterion. (2004, 288)

4.12 Poise

This extra-linguistic criterion was devised by Bühler who defined it as “ease and dignity of manner, expressing a certain amount of (mental and physical) stability”. It ranked low in her study. (1986, 232). Later a similar criterion appeared in the study of Collados Aís called *professionalism* (1998/2002, 333) and of Pradas Macías as *impression of professionalism*. Pradas Macías also lists *impression of reliability* as a criterion. (2006, 35) Bühler used the criterion in an expectation survey, Pradas Macías in a performance assessment and Collados Aís in both.

4.13 Pleasant appearance

The extra-linguistic criterion was used by Bühler (1986, 232). None of the subsequent researchers mentioned in this paper decided to adopt it. *Pleasant appearance* ranked low with 6 respondents rated it as “irrelevant”. (Bühler 1986, 232) Here it was used in an expectation survey. Provided that it would be a video recording, this criterion could also be ranked in an actual performance assessment.

4.14 Reliability

Another extra-linguistic criterion from Bühler that was not adopted later. The criterion was used in her expectation study and it ranked as "highly important". (1986, 232) It would not be possible to judge reliability of the interpreter just from a recorded performance, so it seems that this criterion cannot be judged in performance assessments.

4.15 Ability to work in a team

This extra-linguistic criterion from Bühler's study, where it ranked as "important" and "highly important" (1986, 233), was adopted by Chiaro and Nocella. It was rated the third most important extra-linguistic criterion in their web survey. (2004, 288) *Ability to work in a team* can be ranked in expectation surveys. It is hard to imagine a scenario where it could be reviewed in a particular performance assessment.

4.16 Positive feedback from delegates

A criterion devised by Bühler which was not adopted by others. (1986, 233) The reason for this may be that for most this would constitute a survey of its own.

4.17 Intonation

Intonation was a criterion in the studies of Chin Ng (1992, 38), Collados Aís (1998/2002, 333), Čeňková (1998, 164), Pradas Macías (2006, 33), and AIIC (2008). It was rated as "important" in the AIIC study (2008) and as a parameter with "moderate influence" in Pradas Macías' survey. (2006, 34) Collados Aís noted that "the monotonous intonation was definitely detected by the users". (1998/2002, 335) *Intonation* can be considered to be one of the "superficial" criteria mentioned earlier that users who do not understand the source language of interpretation and thus cannot compare the interpretation with the original use to judge a given performance. It was in the study of Collados Aís where users of interpretation (who could not compare the target text with the source text) judged the performances with mistakes in them better than those without only because the interpretations without mistakes were delivered with monotonous intonation. (1998/2002, 335) Čeňková recommends interpreters to produce a lively delivery which should reflect the mood and the tone of the speaker. (1998, 167) *Intonation* can be ranked in expectation surveys as well as assessed in a performance review by both fellow interpreters and by users who do not speak the source language.

4.18 Synchronicity

This criterion was used in the study of the AIIC (2008) where it was rated as "important" with the participants noting that it is important in certain kinds of speeches, for example, those containing punch lines or numerical data. They also stated that it can be of different importance in different parts of the speech—such as when there is a joke made. (AIIC 2008) This criterion can also be called *ear voice span*, *décalage* or *lagging behind the speaker*. Čeňková calls it *lagging behind the speaker* and considers it to be a negative influence on the performance. She believes that the lag should not exceed one logical unit and that the interpreter should end the delivery at the same time when the speakers ends his speech. (1998, 167) *Synchronicity* can be rated in an expectation study as it happened in this case or it can be evaluated in a performance assessment by either a fellow interpreter or a user who speaks the source language. It can only be used as a criterion when talking about simultaneous interpreting.

4.19 Correct use of equipment

It was a criterion in the survey of Foster which dealt with the difference between "stated and actual quality criteria". (2014, 87) Moser-Mercer talks about *microphone technique* in her paper on evaluation of students who could potentially become interpreters. (1985, 99) Both of these studies had to do with student interpreters which could suggest that professional interpreters should already have a handle on how to use the microphone and so maybe there is no need for it as a criterion in quality assessment.

4.20 Conduct

Can be considered a broader term which can encompass other criteria such as *ability to work in a team, preparation of conference documents, correct use of equipment, poise, professionality.* Foster uses it as one criterion explaining that it stands for: “professional interaction with colleagues, way with which documents and information are dealt”. (2014, 94) In her survey, Čeňková concludes that the interpreter should avoid any kind unprofessional behavior. She names laughter, rustling of papers, whispering, coughing, hums and has and other noises such as from cups and cutlery as examples. (1998, 167)

4.21 Diction

On top of the criteria concerning the voice quality of the interpreter (*accent, intonation, pleasant voice*), Collados Aís (1998/2002, 333) and Pradas Macías (2006, 35) list *diction.* It can either be used as a single criterion or as part of a different one, for example, *pleasant voice.*

4.22 Impartiality

Garber and Mauffette-Leenders used this in their study about obtaining feedback from non-English speakers. They used questions such as “Do you think the interpreter was taking your side in the interpretation?”, “Do you think the interpreter was taking the side of the English speaker in the interpretation?”, or “Do you think the interpreter was working for both sides equally?” in their questionnaire. (1997, 140) *Impartiality* can be rated in an expectation survey and assessed in a performance review by both end users and colleagues.

4.23 Other criteria

Here are some criteria usually only a few authors use in their studies. One of them is *concentration* which was used in the study of Chiaro and Nocella where it was ranked as one of the most important extra-linguistic criteria. (2004, 288) It was also used in the study of Foster but not as a separate criterion, it is mentioned under the category of *interpreting skills.* (2014, 94) Chiaro and Nocella go out and say that *concentration* is a feature essential for interpreting (2004, 289). One might argue that if it is the very prerequisite for interpreting, it should not be used for the purposes of assessing quality. Should we still decide to use concentration as a criterion, it can be used in an expectation study but it would be hard to measure it in any way in a performance assessment.

Another criterion that cannot be fully measured in performance assessments is *stress*. We cannot really measure the actual level of stress the interpreter is under from a given performance. What we can measure or observe is the ability to cope with certain stressful situations. Moser-Mercer used this when evaluating potential students of simultaneous interpretation where she was observing whether they could cope with increased difficulty in exercises with increased input rates, dual-task training, and text material with low redundancy. (1985, 98) Chiaro and Nocella's criterion which they used in their expectations study was *absence of stress*. (2004, 288)

Another extra-linguistic criteria from Chiaro and Nocella's web survey are *physical well-being, mnemonic skills,* and *encyclopedic knowledge*. (2004, 288) This survey was addressed to professional interpreters and it was an expectations survey. It would be hard to implement these three criteria plus *absence of stress* into a performance review. Only with *absence of stress* could there be something measurable such as "signs of stress" in a performance assessment. All of these criteria are extra-linguistic.

5. Conclusion

In trying to find the answer to the question of “Who should assess interpreting?” I found out there is no one solution. Instead of trying to make users of interpreting services assess something they do not understand or forcing fellow interpreters to try to put themselves in the shoes of the users, I believe the answer lies in making different agent assess different aspects of interpreting. Out of the criteria listed, researchers can choose which to include in their survey and decide which assessor is most suitable for each criterion.

The ultimate goal we are trying to reach here, the “holy grail” of interpreting quality assessment, would be sort of a grade which we can attribute to a given performance but that is a distant goal and the question is whether also a desirable one. The answer to that lies in whether we are leaning towards considering an interpreting performance almost a work of art or simply a service that can be rated. I believe that since there are so many settings interpreting can take place in and so many purposes it can serve, it is not appropriate to just attribute the stamp of a “good” or “bad” performance. It is, however, required in interpreting training and there it makes all the sense.

I have decided not to include financial aspects in the list of criteria as I feel that when a potential employer is deciding to hire an interpreter, they should have the possibility of weighing the quality on one side and price on the other. They are like the two sides of a scale and the employer should try to find a balance. As the main focus of the thesis was on assessment criteria, topics such as analysis of results and wording of the different items in a questionnaire were not covered properly and they need further research. Another area that still needs to be covered is the size of our sample. While it is desired for expectation surveys to have a large number of respondents, maybe the number needed for a performance review is much lower? Maybe two graders would suffice? If I could ask interpreting researchers one thing, it would be to include the survey design when they report the results of their research. It would make the job of others interested in quality assessment much easier. A major drawback of several surveys in this field is in the phase before designing the survey itself. With some of them, it is not clear whether they are focusing on finding out expectations of the different agents in the process of interpreting or whether they are assessing an actual performance. All of that should take place in the very first stage of our research. As interpreting quality is a very popular line of research, I believe that these downsides will be rectified.

6. Appendix 1: Table of Criteria

	WHO	PERFORMANCE OR EXPECTATION	P	TYPE OF INTERPRETING	SOURCE
pleasant voice	**UI**	PE	UI	CS	Bühler (1986, 232), Kurz (1993, 15), Collados Aís (1998/2002, 331), Čeňková (1998, 164), Chiaro and Nocella (2004, 287), Pradas Macías (2006, 33), AIIC (2008)
accent	**UI**	PE	UI	CS	Bühler (1986, 232), Kurz (1993, 15), Chin Ng (1992, 38), Collados Aís (1998/2002, 331), Čeňková (1998, 164), Chiaro and Nocella (2004, 287), Pradas Macías (2006, 33), AIIC (2008)
pleasant appearance	**UI**	PE	UI	CS	Bühler (1986, 232)
professionalism, poise	UI	PE	UI	CS	Bühler (1986, 232), Collados Aís (1998/2002, 333), Pradas Macías (2006, 35)
intonation	UI	PE	UI	CS	Chin Ng (1992, 38), Collados Aís (1998/2002, 333), Čeňková (1998, 164), Pradas Macías (2006, 33), AIIC (2008).
style	UI	PE	UI	CS	Bühler (1986, 232), Chin Ng (1992, 38), Collados Aís (1998/2002, 331), Chiaro and Nocella (2004, 287), Pradas Macías (2006, 33), AIIC (2008)
diction	UI	PE	UI	CS	Collados Aís (1998/2002, 333), Pradas Macías (2006, 35)
fluency	UI	PE	UI	CS	Bühler (1986, 232), Kurz (1993, 15), Collados Aís (1998/2002, 331), Čeňková (1998, 164), Chiaro and Nocella (2004, 287), Pradas Macías (2006, 33), AIIC (2008)
correct grammatical usage	UI	PE	UI	CS	Bühler (1986, 232), Chin Ng (1992, 38), Kurz (1993, 15), Čeňková (1998, 164), Chiaro and Nocella (2004, 287), AIIC (2008)
reliability	UI	E		CS	Bühler (1986, 232)
correct terminology	IU	PE	UI	CS	Bühler (1986, 232), Kurz (1993, 15), Collados Aís (1998/2002, 331), Čeňková (1998, 164), Chiaro and Nocella (2004, 287), Pradas Macías (2006, 33), AIIC (2008)
logical cohesion	IU	PE	UI	CS	Bühler (1986, 232), Kurz (1993, 15), Collados Aís (1998/2002, 331), Čeňková (1998, 164), Chiaro and Nocella (2004, 287), Pradas Macías (2006, 33), AIIC (2008)
ability to work in a team	IU	E		CS	Bühler (1986, 233), Chiaro and Nocella (2004, 288)
impartiality	IU	PE	UI	CS	Garber and Mauffette-Leenders (1997, 140)
preparation of documents	IU	E		CS	Bühler (1986, 232), Chiaro and Nocella (2004, 288)
endurance	IU	E (P)	(UI)	CS	Bühler (1986, 232), Moser-Merce (1985, 98), Chiaro and Nocella (2004, 288)
correct use of equipment	IU	PE	UI	CS	Moser-Mercer (1985, 99), Foster (2014, 87)
synchronicity	**IU**	PE	I	S	AIIC (2008), Čeňková (1998, 167)
completeness of interpretation	**IU**	PE	I	CS	Bühler (1986, 232), Kurz (1993, 15), Collados Aís (1998/2002, 331), Čeňková (1998, 164), Chiaro and Nocella (2004, 287), Pradas Macías (2006, 33), AIIC (2008)
sense consistency	**IU**	PE	I	CS	Bühler (1986, 232), Kurz (1993, 15), Collados Aís (1998/2002, 331), Čeňková (1998, 164), Chiaro and Nocella (2004, 287), Pradas Macías (2006, 33), AIIC (2008)

7. Appendix 2: Table Legend

WHO (horizontal)	suggested assessor
P (horizontal)	who can rate in a performance assessment
U	user
I	interpreter or someone speaking both languages
E	can be used in an expectations survey
P	can be used in a performance assessment
C	consecutive interpreting
S	simultaneous interpreting
	criteria better suited for the user
	criteria better suited for the interpreter (or someone speaking both languages)

Shrnutí

Tato bakalářská práce se zabývá hodnocením kvality tlumočení. Snahou bylo se soustředit jak na simultánní tlumočení, tak na konsekutivní, ale zdroje byly především z oblasti simultánního konferenčního tlumočení. V první kapitole se snažím vysvětlit, proč jsem si vybrala právě hodnocení kvality tlumočení jako téma, a proč je potřeba se jím vůbec zabývat. Potřeba hodnocení kvality tlumočení vyplývá mimo jiné z požadavků dnešní konzumní společnosti. Popisuji myšlenkové pochody, které se odehrávají v první fázi výzkumu v oblasti hodnocení kvality tlumočení, a vyjmenovávám rozhodnutí, která musí badatel učinit, aby vybral správné nástroje hodnocení.

Ve druhé kapitole se pokouším odpovědět na otázku, kdo by měl tlumočení hodnotit. Jestli jsou to kolegové-tlumočníci nebo koneční uživatelé hotového produktu. Ukazuje se, že odpověď není jednoduchá a že neexistuje jedna odpověď.

Třetí kapitola provází procesem tvorby dotazníků pro potřeby hodnocení tlumočení. Navrhuji zde jakýsi návod, jak si vybrat druh výzkumu, který chceme provést. Dále vyjmenovávám metody, jakými lze vybrat respondenty pro náš dotazník. V této kapitole také s pomocí příručky od Sandry Hale a Jeminy Napier popisuji, jak by měl takový dotazník vypadat a jak bychom ho měli distribuovat. Velice důležitou částí této kapitoly je vyjmenování různých druhů analýzy výsledků našeho výzkumu.

Čtvrtá kapitola se zabývá samotnými kritérii hodnocení a tím, kdo podle nich může hodnotit. Docházím k závěru, že si nelze vybrat pouze hodnocení od tlumočníku nebo pouze od publika, ale že ideálním hodnocením by byla syntéza těchto dvou hodnocení.

V závěru uvádím, že společným cílem vývoje nástroje k hodnocení tlumočení je možnost udělení jakési „známky“ jednotlivým výkonům. Otázka zní, jestli je takový nástroj žádoucí. Zmiňuji zde také, proč jsem se rozhodla nezahrnout finanční aspekty do svého seznamu kritérií hodnocení. Dále navrhuji, v jakých oblastech problematiky je potřeba dalšího výzkumu.

Cílem práce je poskytnout ostatním výzkumníkům seznam kritérií, která mohou použít pro potřeby vlastního výzkumu a navrhnout jakýsi návod, jak vybrat správný nástroj hodnocení kvality tlumočení.

Bibliography

AIIC. 2008. “Survey on quality and role: conference interpreters’ expectations and self-perceptions.” Accessed December 6. http://aiic.net/page/3405/survey-on-quality-and-role-conference-interpreters-expectations-and-self-perceptions/lang/1.

Barik, Henri C. 1971. “A Description of Various Types of Omissions, Additions and Errors of Translation Encountered in Simultaneous Interpretation.” *Meta: Translators’ Journal* 16(4): 199-210. doi: 10.7202/001972ar.

Bühler, Hildegund. 1986. “Linguistic (semantic) and extra-linguistic (pragmatic) criteria for the evaluation of conference interpretation and interpreters.” *Multilingua - Journal of Cross-Cultural and Interlanguage Communication* 5(4):231-235. doi:10.1515/mult.1986.5.4.231.

Čeňková, Ivana. 1998. “Quality of Interpreting - a Binding or a Liberating Factor?” In *Translators’ Strategies and Creativity: Selected Papers from the 9th International Conference on Translation and Interpreting, Prague, September 1995*. Edited by Ann Beylard-Ozeroff, Jana Králová and Barbara Moser-Mercer, 163-170. Amsterdam/Philadelphia: John Benjamins.

Chiaro, Delia and Giuseppe Nocella. 2004. “Interpreters’ Perception of Linguistic and Non-Linguistic Factors Affecting Quality: A Survey through the World Wide Web.” *Meta: Translators’ Journal* 49(2): 278-293. doi:10.7202/009351ar.

Chin Ng, Bee. 1992. “End Users’ Subjective Reaction to the Performance of Student Interpreters.” *The Interpreters’ Newsletter Special Issue* 1: 35-41.

Collados Aís, Ángela and Macarena Pradas Macías and Elisabeth Stévaux and Olalla García Becerra. 2007. *La evaluación de la calidad en interpretación simultánea: parámetros de incidencia*. Granada: Comares.

Collados Aís, Ángela. 1998/2002. "Quality assessment in simultaneous interpreting: The importance of nonverbal communication." In *The Interpreting Studies Reader*. Edited by Franz Pöchhacker and Miriam Shlesinger, 327-336. London/New York: Routledge.

Dörnyei, Zoltán. 2007. *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.

Foster, Lenelle. 2014. "Quality-assessment expectations and quality-assessment reality in educational interpreting: An exploratory case study." *Stellenbosch Papers in Linguistics Plus* 43:87-102. doi:10.5842/43-0-207.

Garber, Nathan and Louise A. Mauffette-Leenders. 1997. "Obtaining feedback from non-English speakers." In *The Critical Link: Interpreters in the Community: Papers from the 1st international conference on interpreting in legal, health and social service settings, Geneva Park, Canada, 1–4 June 1995*. Edited by Silvana E. Carr, Roda P. Roberts and Aideen Dufour, 131-143. Amsterdam/Philadelphia, John Benjamins.

Garzone, Giuliana. 2002. "Quality and norms in interpretation". In *Interpreting in the 21st century: Challenges and opportunities*: *Selected papers from the 1st Forlì Conference on Interpreting Studies, 9-11 November 2000*. Edited by Giuliana Garzone and Maurizio Viezzi, 107–120. Amsterdam/Philadelphia: John Benjamins.

Gile, Daniel. 1994. "Methodological aspects of interpretation and translation research". In *Bridging the Gap: Empirical Research on Simultaneous Interpretation*. Edited by Sylvie Lambert and Barbara Moser-Mercer, 39-56. Amsterdam/Philadelphia: John Benjamins.

Hale, Sandra and Jemina Napier. 2013. *Research Methods in Interpreting*. London/New York: Bloomsbury.

Kalina, Sylvia. 2005. "Quality Assurance for Interpreting Processes." *Meta: Translators' Journal* 50(2): 768-784. doi:10.7202/011017ar.

Kellett Bidoli, Cynthia Jane. 1996. "Quality Assessment in Conference Interpreting: an Overview." *Miscellanea* 4: 105-145.

Kurz, Ingrid. 1993. "Conference Interpretation: Expectations of Different User Groups." *Interpreters' Newsletter* 5: 13-21.

Kurz, Ingrid. 2001. "Conference Interpreting: Quality in the Ears of the User." *Meta: Translators' Journal* 46(2): 394-409. doi:10.7202/003364ar.

Moser-Mercer, Barbara. 1985. "Screening Potential Interpreters." *Meta: Translators' Journal* 30(1): 97-100. doi:10.7202/003631ar.

Moser-Mercer, Barbara. 1996. "Quality in interpreting: some methodological issues." *The Interpreters' Newsletter* 7:43-55.

Pöchhacker, Franz. 1994. "Quality assurance in simultaneous interpreting." In *Teaching Translation and Interpreting 2: Insights, aims and visions. Papers from the Second Language International Conference Elsinore, 1993.* Edited by Cay Dollerup and Annette Lindegaard, 233-242. Amsterdam/Philadelphia: John Benjamins.

Pöchhacker, Franz. 2001. "Quality Assessment in Conference and Community Interpreting." *Meta: Translators' Journal* 46(2): 410-425. doi:10.7202/003847ar.

Pöchhacker, Franz. 2005. "Quality research revisited." *The Interpreters' Newsletter* 13: 143–166.

Pradas Macías, Macarena. 2006. "Probing quality criteria in simultaneous interpreting." *Interpreting* 8(1):25-43. doi:10.1075/intp.8.1.03pra.

Seleskovitch Danica. 1986. "Comment: Who should assess an interpreter's performance?" *Multilingua - Journal of Cross-Cultural and Interlanguage Communication* 5 (4): 236. doi:10.1515/mult.1986.5.4.236.

Shlesinger, Miriam and Franz Pöchhacker. 2002. *The Interpreting Studies Reader*. London/New York: Routledge.

Shlesinger, Miriam. 1997. "Quality in Simultaneous Interpreting." In *Conference Interpreting: Current Trends in Research: Proceedings of the International Conference on Interpreting: What do we know and how?* Edited by Yves Gambier, Daniel Gile and Christopher Taylor, 123-131. Amsterdam/Philadelphia: John Benjamins.

Tiselius, Elisabet. 2005. "The Work Shows the Workman – An Investigation of Two Methods for Measuring Interpreting Performance and Studying Expertise in Interpreting." M.A. thesis, Stockholm University.

Tiselius, Elisabet. Interpretings.net. 2013. "How to assess interpreting." Accessed December 6. http://interpretings.net/2013/03/20/how-to-assess-interpreting/.

Anotace

Autor: Karolína Tomanová
Katedra: Katedra anglistiky a amerikanistiky, FF UPOL
Název česky: Kritéria hodnocení kvality tlumočení
Název anglicky: Criteria of Quality Assessment in Interpreting
Vedoucí práce: PhDr. Veronika Prágerová, Ph.D.
Počet stran: 31
Počet znaků: 60 517

Anotace:

Práce se věnuje hodnocení kvality tlumočení. Zaměřuje se jak na hodnocení tlumočení jako konečného produktu, tak na zjišťování požadavků, jaké si nejrůznější účastníci procesu tlumočení na tlumočení kladou. Je zde snaha zodpovědět otázku, kdo by měl tlumočení hodnotit. Třetí kapitola provází procesem vytváření dotazníkového šetření pro potřeby hodnocení tlumočení. Závěrečná kapitola nabízí seznam kritérií, podle kterých lze tlumočení hodnotit.

Klíčová slova: kvalita tlumočení, kritéria hodnocení kvality tlumočení, sestavování dotazníků pro účely výzkumu tlumočení

Abstract:

Interpreting quality assessment is the topic of this thesis. It focuses both on assessing an interpreting performance and on finding out expectations of the different agents in the process of interpreting. We try to answer the question of who should assess interpreting. The third chapter describes the process of designing a questionnaire for the purposes of assessing interpreting quality. The last chapter offers a list of criteria which can be used when assessing interpreting.

Key words: interpreting quality, criteria of quality assessment, questionnaire design in interpreting