Language Testing and Evaluation 29

Dina Tsagari / Roelof van Deemter
(eds.)

Assessment Issues in Language Translation and Interpreting

PETER LANG
EDITION

Assessment Issues in Language Translation and Interpreting

Language Testing and Evaluation

Series editors: Rüdiger Grotjahn
and Günther Sigott

Volume 29

PETER LANG
EDITION

Dina Tsagari / Roelof van Deemter
(eds.)

Assessment Issues in Language Translation and Interpreting

PETER LANG
EDITION

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the internet at http://dnb.d-nb.de.

Cover Design:
© Olaf Glöckler, Atelier Platen, Friedberg

Library of Congress Cataloging-in-Publication Data

Assessment issues in language translation and interpreting /
Dina Tsagari, Roelof van Deemter (eds.)
pages cm. — (Language testing and evaluation ; Volume 29)
ISBN 978-3-631-63603-9
1. Translating services. 2. Translating and interpreting.
3. Translating services—Evaluation. 4. Translating and interpreting—Evaluation. I. Tsagari, Dina.
P306.94.A88 2013
418'.020287—dc23

2013006368

ISSN 1612-815X
ISBN 978-3-631-63603-9

To our families, friends and colleagues

Table of Contents

Foreword 9
Claudia V. Angelelli

Part I
General Issues in Assessing Language Translation and Interpreting 13

How Do We Assess Students in the Interpreting Examinations? 15
Fred S. Wu

Assessing Interpreter Aptitude in a Variety of Languages 35
Hanne Skaaden

Unpacking Delivery Criteria in Interpreting Quality Assessment 51
Emilia Iglesias Fernández

Rethinking Bifurcated Testing Models in the Court
Interpreter Certification Process 67
Melissa Wallace

Assessing the Impact of Text Length on Consecutive Interpreting 85
Jungyoon Choi

Translation versus Language Errors in Translation Evaluation 97
Tomás Conde

Part II
Applications of Assessing Language Translation and Interpreting
in Local System 113

The System of Authorizing Translators in Finland 115
Leena Salmi and Ari Penttilä

Translation Competence and the Practices of Translation Quality Assessment in Turkey 131
Nilgun Dungan

Evaluating Assessment Practices at the MCI in Cyprus 145
Georgios Floros

Design and Analysis of Taiwan's Interpretation Certification Examination 163
Minhua Liu

Certification of Social Interpreters in Flanders, Belgium: Assessment and Politics 179
Britt Roels

Foreword

Claudia V. Angelelli[1]
San Diego University

Translation and Interpreting in Lan Iguage Assessment contributes to the body of knowledge on testing, measurement and assessment in Translation and Interpreting Studies (TIS). Conceptualizing translation and/or interpreting as constructs and stating their sub-components, as well as finding out how well someone translates/interprets is no simple undertaking. In TIS, issues of measurement and assessment (beyond exploring the construct of quality) have begun to gain researchers' attention only recently. In 2009 we had the first volume on this topic (Angelelli & Jacobson, 2009) that focused on issues in Translation and Interpreting (T&I) assessment across languages and settings with a focus on both processes and products. This was followed by a special issue of Interpreting (Shlesinger & Pöchhacker, 2011) focusing on aptitude testing. Now this collection edited by Dina Tsagari and Roelof van Deemter captures a broad range of issues and themes. Covering a variety of languages and areas of the world as well as various professional and instructional settings (e.g. graduate, undergraduate and certificate programs and standalone courses) this volume raises important questions in an area currently under scrutiny: the measurement and assessment of translators and interpreters and the interplay of language, translation and interpreting. It is organized in two parts. Part I contains six chapters that present general issues in assessing translation and interpreting. Part II presents five chapters that discuss applications of translation and interpreting assessment in local systems.

Part I opens up with Fred Wu's contribution "How Do We Assess Students in the Interpreting Examinations?" He addresses the concerns raised on the consistency that professional interpreters may or may not exhibit when asked to assess student-interpreters' performances. Wu presents findings of an experimental pilot study designed to ascertain the reported fuzziness in the use of assessment criteria and inconsistent judgment in interpreting examinations that may be exhibited by judges. Addressing admission tests of undergraduate Norwegian students, Hanne Skaaden presents results of qualitative analyses conducted across 50 languages between 2007 and 2011 in "Assessing Interpreter Aptitude in a variety of Languages". Skaaden shows the importance of a high level of bilingual proficiency in order to undertake studies in interpreting and to perform as professional interpreters.

1 cangelel@mail.sdsu.edu

Stressing the notion of criteria, Emilia Iglesias Fernández argues for "Unpacking Delivery Criteria in Interpreting Quality Assessment." The author states that the overly simplified view of language and speech, which permeates the culture of assessment in interpreting, falls short of capturing the interpreting phenomenon. In addition she argues that the refinement of presentation-related categories (such as intonation, diction, pleasant voice) for assessment of interpreting quality is essential to avoid unnecessary overlapping. This is particularly important for preserving inter-item consistency.

In her chapter entitled "Rethinking Bifurcated Testing Models in the Court Interpreter Certification Process", Melissa Wallace explores whether or not success in one mode of interpreting (e.g. simultaneous) on the US Consortium for Language Access in the Courts' oral certification exam could potentially predict successful performance in the other two modes (e.g. consecutive and sight translation). In addition, her work explores examining individual modes of interpreting as potential predictors of success of the entire oral court certification exam, as well as contemplating the potential for utilizing such information in the context of interpreter certification testing.

Jungyoon Choi's contribution is entitled "Assessing the Impact of Text Length on Consecutive Interpreting". It discusses how source text variables, such as text length, potentially could influence interpreters' performance. Based upon the hypothesis that longer texts are likely to require more concentration and endurance from interpreters than shorter ones, the author presents the results of an experiment conducted on future interpreters at the graduate level.

In the last chapter of Part I entitled "Translation versus Language Errors in Translation Evaluation", Tomás Conde compares the number of issues in translation versus language (errors and good decisions) as graded by two groups: one composed of extra academic evaluators (professional translators and potential addressees of translations) and one of academic evaluators (teachers and students of translations). His analysis shows that errors as well as good decisions in translations analyzed do have different relevance and incidence rates for the type of evaluators studied.

Part II opens up with Leena Salmi and Ari Penttilä's contribution entitled "The System of Authorizing Translators in Finland." In this chapter, after a historical account, the authors provide a detailed description of the process used in Finland to certify translators as updated in 2008. The chapter reports on four years of data of exams (2008–2011). Discussing issues related to translation, Nilgun Dungan's "Translation Competence and the Practices of Translation Quality Assessment in Turkey" argues for thinking about assessment beyond translation equivalence and for the development of objective sets of criteria to judge translation quality. This chapter discusses how key competencies in translation (even as limitedly

defined as they currently are) are closely related to enhancing employability of the graduates.

The third chapter in Part II is Georgios Floros' "Evaluating Assessment Practices at the MCI in Cyprus". The author discusses assessment practices in a graduate program of conference interpreting in Cypress. Building on three years of data from the Masters on Conference Interpreting, Floros' work focuses on problems related to measuring general background knowledge and booth manners, as well as problems that arise from the use of a general rating scale.

In her chapter entitled "Design and Analysis of Taiwan's Interpretation Certification Examination" Minhua Liu describes analyses conducted to study the Taiwanese ECTICE exams as a valid and full-fledged examination for professionals. In dialogue with Liu's contribution Britt Roels' work closes Part II. Her chapter entitled " Certification of Social Interpreters in Flanders, Belgium: Assessment and Politics" outlines the processes involved in the development phases of an objective, reliable and valid assessment procedure for the certification of social interpreters (SI) in Flanders. In addition, this chapter highlights the political and historical background of the Flemish SI sector and its uncertain future.

In sum, as we will see from the contributions to this volume, discussions on valid and reliable ways of measuring translation and interpreting processes and products are as essential as they are complex. As we continue to transfer information across languages and cultures, and to engage in cross-cultural/linguistic interactions the need for competent professionals as well as for quality translation and interpreting is ever more pressing. Using valid and reliable ways of measuring performance or quality is therefore essential.

The issues discussed in this volume continue to shape approaches to measurement in translation and interpreting. The questions raised by contributors merit the attention of key players in TIS. The field of measurement and assessment in translation and interpreting is growing but there is still much work to be done. This volume is certainly a step in the right direction.

References

Angelelli, C V. & H. E. Jacobson (eds.) 2009. *Testing and Assessment in Translation and Interpreting Studies*. ATA Scholarly Monograph Series. Amsterdam: John Benjamins.

Shlesinger, M. & F. Pöchhacker (eds.) (2011). *Aptitude for Interpreting.* (Special issue) *Interpreting* 13(1).

Part I
General Issues in Assessing Language Translation and Interpreting

How Do We Assess Students in the Interpreting Examinations?

Fred S. Wu[1]
Newcastle University, United Kingdom

The field of assessment in interpreter training is under-researched, though trainers and researchers have pointed out urgent issues to be addressed in this field. Among them, issues surrounding test validity and reliability are most in need of clarification. This study tackles this subject by exploring what examiners really pay attention to when assessing student interpreters, and verifies the concerns about judgement consistency in interpreting examinations. Based on the study findings, a conceptual model is proposed as a framework for further explorations into the relationships and interactions between the elements of interpreting assessment.

Key words: simultaneous interpreting, assessment criteria, examiner behaviour, test reliability.

1. Introduction

With the ever increasing international demand for multilingual communication, there has been a boom in demand for training conference interpreters. To ensure the quality of interpreter training, assessment is crucial. In interpreter education, assessment usually refers to evaluating students' learning outcomes, identifying their strengths and weaknesses, which normally involves assigning a mark or a grade to the students' performances.

There are problems, however, when interpreting assessment methods are scrutinised by using fundamental concepts of assessment, like *validity* and *reliability,* from more established disciplines, such as language testing and educational assessment. The design and administration of interpreting examinations in many higher education institutions still heavily rely on the professional experience of staff, often with no basis in empirical studies for test items and procedures (Liu, Chang & Wu, 2008, p. 35), and test designs have been described as "intuitive" (Campbell & Hale, 2003, p. 211). This lack of empirical base has raised concerns about the reliability and validity of interpreting examinations because test constructs and assessment criteria arguably require clear definitions and descriptions.

Research into interpreting assessment, however, is still at an exploratory stage, and many important concepts and instruments, such as test constructs and assess-

1 fred.wu@newcastle.ac.uk

ment criteria, are still underdeveloped. When discussing these concerns, Angelelli and Jacobson (2009, p. 21) pointed out that

> knowing a situation intimately and defining it clearly for testing purposes are two very distinct things. Definition of a target construct often takes a particular kind of expertise that is different from the expertise of a practitioner. The practitioner is in the midst of the target situation and sometimes fails to notice aspects of the situation merely because they are taken for granted.

Sawyer urged that "if validation is a rhetorical art, it is one at which the community of interpreter educators should excel" (2004, p. 235). After all, if test designers and examiners "are unable to express their subjective judgments by objectively measurable standards" (Kalina, 2005, p. 768), it will be difficult for interpreting examinations to be truly reliable.

Within this context, this chapter reports an attempt to explore and better understand the various dimensions and test constructs of interpreting examinations, and proposes a conceptual model for describing them.

2. The concerns and a starting point to address them

Serious concerns have been raised about how professionals in the interpreting field can judge interpreting performances consistently. Sawyer observed that how interpreter examiners applied assessment criteria was "fuzzy" (2004, p. 185), and that examiners' expertise did not necessarily translate into a high degree of agreement between professional judgements: hence more systematic studies of assessment procedures were urgently needed (ibid, pp. 187–189).

Performance assessment has long been criticised as unreliable and in need of systematic study (Campbell & Hale, 2003, p. 212) and concerns about professional judgment are mainly due to its subjective nature (Messick, 1989, p. 91). Therefore, proper test instruments and procedures are usually required to facilitate a sound and reliable judgement and to report test results by combining examiners' qualitative and quantitative decisions (Pollitt & Murray, 1996, p. 74). However, any well thought-out examination criteria, procedures and test instruments will be of little value in test reliability, and therefore validity, if examiners do not use them consistently or if the design of the instrument itself makes it hard to use them consistently.

Studies of language testing also identify examiners themselves as a source of measurement error (Alderson, Clapham and Wall, 1995; Bachman, Lynch & Mason, 1995; Fulcher, 2003; Lumley & McNamara, 1993; Luoma, 2004). Such error can subtly influence the results of performance-based assessments, making

assessment procedures unreliable and threatening test validity (Eckes, 2005, p. 197). Test instruments, such as rating scales with specific assessment criteria, and examiner trainings are often used to help reduce subjectivity in assessment and increase consistency between examiners.

In language speaking tests, however, researchers pointed out that many rating scale descriptors were created to look consistent with little empirical basis. They suggested that rating scales should match what the examiners actually perceive in the performances they have to grade, and argued that the scale development should start from studying "the perceptions of proficiency by raters in the act of judging proficiency" (Fulcher, 2003; Pollitt & Murray, 1996, p. 76). These experiences in language testing provide valuable lessons for the study on the interpreting assessment.

Taking the background and rationale above, this study was conducted to explore and understand how individual examiners perceive the interpreting performances in a simultaneous interpreting examination, and how they make the judgments. The study method and its main findings are summarised below, and based on the findings, a conceptual model is proposed to illustrate the relationships between the various elements in a typical interpreting examination.

3. Research method

A simulated examination of simultaneous interpreting was conducted for the study. However, as a consensus remains to be established on an empirical-based standard assessment procedure and test instrument for the interpreting examinations (Liu et al., 2008, p. 35), using a potentially flawed rating scale in a study that employs psychometric method will impose higher limitations in generalising the research findings (Caban, 2003, p. 34). Therefore, it would not be ideal to base a research study on a rating scale and an examiner training session of the interpreting examinations that are both intuitively designed, which may risk the validity of the study. An alternative research approach is needed.

Studies on the rater-related issues in language testing also went through "a phase of exploration" (Lumley & McNamara, 1993, p. 5), and encountered some problems that could not be addressed solely by using the quantitative-oriented psychometric research method (Upshur & Turner, 1999, pp. 103–107). Qualitative research approach was suggested to supplement the statistical method because there is almost always a qualitative element present in the process of making judgements; qualitative approaches provide insights into how experts make judgements, which cannot be gained from statistical analysis (Fulcher, 2003, pp. 216–224). Therefore, qualitative data is crucial if the study aim is to explore and

gain insights into how the examiners make judgements in the interpreting examinations.

Integrating both quantitative and qualitative methods in a research project "may provide a better understanding of a phenomenon than if just one method had been used" (Bryman, 2004, pp. 452–464). Pollitt and Murray successfully demonstrated the usefulness of a mixed-method study design to elicit the constructs of the rating scale for speaking test. They employed Thurstone's Method of Paired Comparisons to monitor the examiners' consistency levels, i.e. quantitative approach, which also "facilitated the expression by the judges of the aspects that seemed salient to them", i.e. qualitative approach (Pollitt & Murray, 1996, pp. 74–91). This method is useful for its flexibility that allows the examiners to express their judgements on the examinees' performances, and at the same time for the researchers to systematically record and analyse the study data.

For the purpose of this study on interpreting assessment, another useful aspect of the Paired Comparison method is that it does not require a rating scale, but only requires the examiners to compare items two by two and decide which one is better. Therefore, the design of this study takes a multi-strategy approach by employing both quantitative and qualitative methods. The Method of Paired Comparisons was used to collect quantitative data for monitoring the examiners' judgement consistency levels. While making comparisons, the examiners were also asked to think aloud their judgement and comment on the students' performances. The examiners' comments (qualitative data) were recorded and coded for analysis, extracting key concepts in the examiners' judgement process.

3.1 Study procedures

A pilot study with eight participant examiners was first conducted to ascertain the reported fuzziness in the use of assessment criteria and inconsistent judgement in interpreting examinations; it also verified the usefulness of the proposed research methods (Wu, 2010). Based on the refined study procedures from the pilot study, thirty examiners were recruited to participate in the main study.

In language testing, researchers noticed that the consistency level of judgement was impressive among the non-specialist examiners, i.e. those who had little or no experience of the formal oral assessment of languages (Pollitt & Murray, 1996, p. 88). By contrast, it was noted that there were clear variations in interpreter examiners' professional judgements (Sawyer, 2004, p. 188). These observations of the examiners' judgements prompted this study to include both interpreter and non-interpreter examiners as participants in order to generate contrastive data for analysis. The participant examiners came from three main backgrounds:

- Professional interpreters with substantial experience in SI teaching
- Professional interpreters with little or no experiences in SI teaching
- Professional translators and/or translation teachers with some or no interpreting training

In this study, there are 19 interpreter examiners and 11 non-interpreter examiners, whose working languages are Mandarin Chinese and English, with Mandarin being the first language of all the examiners except one who was based in the UK.

Table 1. Student background information for main study

Student / Code (pseudonyms)		**Course exam mark**	**A Language**	**B Language**
Ally /	**A**	low 50	Chinese	English
Beth /	**B**	mid 60	Chinese	English
Cherry /	**C**	70+	English	Chinese
Daisy /	**D**	mid 50	Chinese	English
Eileen /	**E**	high 50	Chinese	English

Authentic examination recordings (English-into-Chinese simultaneous interpreting) of five postgraduate students were selected for the main study as shown in Table 1. As the study investigates normal assessment behaviour of examiners, not the students themselves, levels of students' interpreting abilities were preselected, ranging from the highest marked performers to the lowest ones according to the marks given in one interpreting examination. It was hoped that a wider range of performance levels would elicit more insights from the examiners when they compared the student performances.

The English source speech in the examination task for study was a three-minute excerption selected from a keynote speech in a business conference. The examination recordings were made in digital video format with the students' Chinese interpretations in the main sound track and the English source speech in the secondary sound track. The participant examiners of this study, therefore, could watch the students performing simultaneous interpreting from the video recordings, and simultaneously monitor both the target and source languages.

It was unlikely to gather all thirty participant examiners under one roof for the study. Therefore, for practical reasons, the examination simulations were conducted with one examiner at a time in each session. Following the same study

procedures, the examiners were asked to compare the students' performances in pairs. Given *n* students, there should be *n(n-1)/2* pairs in total to compare. So with five students, there were *ten* pairs to compare. The results of the paired comparisons, i.e. the number of times a student was judged better, were added up and converted into ranking points; 5 indicates the best performance and 1 is the worst.

Immediately after viewing each pair, the examiners were asked to compare and decide which one was better, and at the same time to think aloud their judgements on the performances, in what way they were better or worse, similar or different and any other relevant comment. The verbal comments were recorded for analysis later. After comparing the ten pairs, the examiners then gave their overall judgement rankings and marks of the five student performances.

4. Study results and discussion

The above study process generated two types of data: (1) the quantitative data, i.e. three sets of ranking points of the five students – paired comparisons (PC), overall judgement (OJ), and the rankings of the final overall marks (OM), and (2) the qualitative data, i.e. the examiners' verbal comments while making the comparisons.

4.1 Quantitative results – examiner reliability

The thirty examiners' judgements on the five student interpreters were evaluated using ANOVA. All three *p* values[2] are less than 0.001, indicating that the five students are highly significantly different in terms of their rankings and marks within each of the three assessment methods. Therefore, we can confidently say that the thirty examiners *as a group* are successful in separating the five students' interpreting performances when using the three assessment methods – though there are wide variations between *individual examiners*, as will be explained shortly.

Figure 1. PC Thurstone scales of interpreting proficiency

Better				*Worse*
Cherry	Beth	Eileen	Daisy	Ally
0.845	0.675	-0.262	-0.502	-0.755

2 PC: $F(4,154) = 37.097$, $p < 0.001$; OJ: $F(4,154) = 42.524$, $p < 0.001$; OM: $F(4,154) = 16.792$, $p < 0.001$

Figure 2. OJ Thurstone scales of interpreting proficiency

Better				*Worse*
Cherry	Beth	Eileen	Daisy	Ally
0.933	0.632	-0.379	-0.443	-0.743

Figure 3. OM Thurstone scales of interpreting proficiency

Better				*Worse*
Cherry	Beth	Eileen	Daisy	Ally
0.837	0.688	-0.377	-0.382	-0.766

**Actual examination marks: Cherry (71), Beth (66), Eileen (58), Daisy (55), Ally (52)*

An alternative way of looking at these data is given by the three Thurston scales (hereafter T scales, Figures 1, 2, 3) which were produced based on the ranking datasets from the three assessment methods. The T scales can be regarded as interpreting proficiency scales to show the five students' relative positions according the thirty examiners' aggregated judgements. On each T scale, the order of the five students is the same, i.e. the thirty examiners *as a group* judged the five students consistently between the three assessment methods. Here, the students' relative positions and distances on the T scales are also a perfect match to the marks the five students received in the actual examination.

The only noticeable difference among the three T scales is the gap between Eileen and Daisy, which appears wider on the PC T scale than on the other two T scales. This variation in the gap may indicate that Eileen and Daisy have a similar level of interpreting proficiency so the examiners put them closer when it comes to more general judgements, such as in the OJ and OM assessment methods. Since examiners were asked to choose a winner, the larger gap on the PC T scale may also result from the fact that examiners had to make a distinction where the two students might otherwise have been considered as similar, if not equal.

In other words, the OM method may be more "accurate" in describing the student interpreters' ability levels in terms of their relative distances. However, it may also be more difficult to maintain a good consistency level of the examination results by using the OM method because examiners may not agree on every detail of the interpreting performances and give the same judgement. This is also shown statistically in Table 2 where the

Table 2. Cronbach's alpha (ICC) for all examiners

All examiners' judgments		Intra-class correlation	95 % confidence interval	
			Lower bound	Upper bound
Paired Comparison	Single measures	0.49	0.24	0.90
	Average measures	0.97	0.90	1.00
Overall Judgment	Single measures	0.52	0.26	0.90
	Average measures	0.97	0.92	1.00
Overall Mark	Single measures	0.41	0.18	0.86
	Average measures	0.95	0.87	0.99

OM method has the lowest score (0.41) of Cronbach's alpha intra-class correlation coefficient, which indicates the reliability level when only a single item is used. The low values of the three single-measures ICC scores (0.49, 0.52, 0.41) suggest poor and unacceptable consistency levels of *individual* examiners' judgements when assessing the students. These statistical results reflect the observed between-examiner fluctuations in this study, which can be illustrated as the ranking point line graph of the thirty examiners in Figure 4.

So far, it appears that statistically the thirty examiners as a group assessed the students with a good consistency level. However, it is impractical to use a thirty-examiner panel in an interpreting examination to achieve a consistent test result. As the single-measures ICC shows that individual examiners are less likely to be reliable, it would be useful to find out what the minimum number of the examiners could be to achieve a reliable level of test result. For example, can a smaller group of examiners of the same background achieve a satisfactory reliability level?

Figure 4. Line graph of paired comparison (PC) rankings

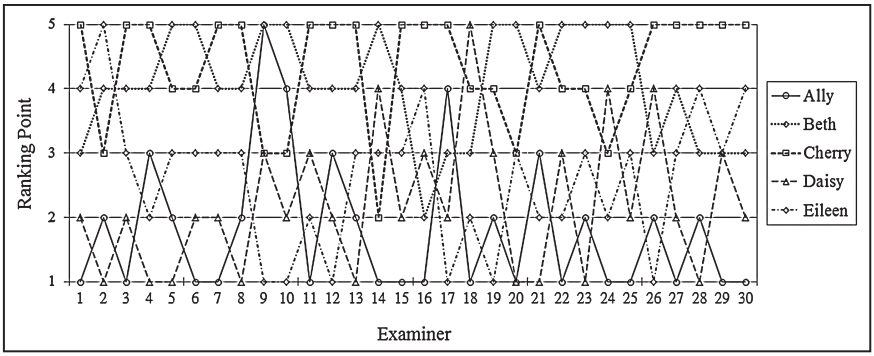


Table 3. Cronbach's alpha (ICC) – according to examiners' background

Examiners / ICC scores	Paired Comparison		Overall Judgment		Overall Mark		Number of Examiners
	Single	Average	Single	Average	Single	Average	
Interpreters	0.43	0.93	0.48	0.95	0.39	0.92	19
Teaching SI	0.46	0.92	0.55	0.94	0.43	0.91	13
Translator	0.72	0.97	0.68	0.96	0.48	0.91	11

Table 3 shows the ICC scores according to the examiners' background. Again, the single-measure ICC scores are generally low, especially in the OM method. This suggests that regardless of the examiners' background, the consistency level would be unacceptable if only one individual examiner is used for marking the interpreting examinations.

The only occasion where the individual examiners can be considered as judging at an acceptable consistency level is the PC method by translator examiners (0.72). This is an interesting comparison to the findings in language testing studies, in which the non-specialist examiners showed "impressive" consistency level of judgement (Pollitt & Murray, 1996, p. 88). Although one could argue that the non-specialist examiners might be consistently wrong, the specialist examiners, in the case of interpreter examiners here, do show lower consistency levels in judgement. Therefore, the logical next step would be to look at the qualitative data, i.e. the examiners' comments on the student interpreters' performances, to understand their use of assessment criteria when making judgements.

4.2 Qualitative results – assessment criteria

In this section, we try to explore and answer two questions: (1) what are the assessment criteria that were actually used by the examiners in this study? and (2) are the ranking patterns of the student interpreters' performances the results of differences in examiners' assessment criteria, or did they use similar criteria but, on this basis, ranked students differently?

The examiners' paired comparison comments were transcribed for line-by-line coding and analysis. Table 4 shows an example of the coding process. When a distinctive idea or concept was identified, the conceptual property was coded by using a key word or phrase. The idea or concept was

Table 4. Example of initial open coding – comparing Ally and Cherry

Examiner's comment	Coded concepts
English translation (by present author): Overall C is much better than A. Her voice is sweet; her pace is steady stable without suddenly picking up speed or slowing down. She seldom has excessive long pauses, and has less meaningless, empty fillers. My impression is that she did pretty well in the first two-thirds of the task. Toward the end she probably was also aware that she did not hear and missed some important numbers. The market share percentage should have been 35 % to 40 %. She said 45 %. Her Chinese sounded very awkward and not fluent here compared with other sentences. This might be because that she was busy remembering the numbers.	**PD:** sweet voice, steady, stable pace, pauses, fillers, fluency **EB:** examiner impression **FAI:** listening comprehension, **FC:** omissions, message weighting, numbers, terminology, **EB:** holistic judgement, examiner speculation PD: Presentation & Delivery EB: Examiner Behaviour FAI: Foundation Ability for Interpreting FC: Fidelity & Completeness

the subjective articulation of the examiners' thinking during the paired comparison judgement. As the aim is to identify the assessment criteria, the coding process focused on any conceptual key words from which inferences can be drawn on how the examiner judged the interpreting performances. In addition, any comments that show how the examiner used the criteria were also coded, i.e. examiner behaviour (EB).

At the end, five categories of assessment criteria were identified: Presentation and Delivery (PD), Fidelity and Completeness (FC), Audience Point of View (APV), Interpreting Skills and Strategies (ISS), and Foundation Abilities for Interpreting (FAI). Each criterion also contains various properties that the examiners distinguish when assessing the student interpreters; this means that they are useful for implementations in a test. These conceptual properties, for example, may be used as a base to operationalize the test constructs of interpreting examinations, formulating descriptors for rating scales to assign marks. Table 5 shows the identified assessment criteria and their main conceptual properties.

Many of the properties in the five assessment criteria were also found to be closely related to one another and difficult to judge separately. It appears that *language competency*, such as listening and speaking skills of the two working languages, permeates all of the assessment criteria, making the judgement work complex to do because each of the five criteria is interrelated in some way. The permeation may be the underlying reason why the use of assessment criteria is fuzzy at the operational level. So the examiners usually resort to holistic judgements, i.e. by impression, leading to some inconsistencies in the examination results.

Table 5. Five identified assessment criteria and their conceptual properties

Assessment criteria	Conceptual properties
Presentation and Delivery	3 aspects: • acoustic • word/phrase • flow of information
Fidelity and Completeness	3 areas: • content accuracy • speaker intention • contextual consistency
Audience Point of View	• to have the confidence in the speaker (via the delivery style of interpretation) • to receive the speaker's message at an acceptable level of faithfulness.
Interpreting Skills and Strategies	• resourcefulness: the ability to use skills and strategies, such as paraphrasing, summarising, skipping, self-correction, background knowledge and anticipation. • multi-tasking: supports using the interpreting skills and strategies. The multi-tasking ability can be observed by looking at the way interpreters manage their Ear-Voice-Span (EVS), or lags.
Foundation Abilities for Interpreting	• listening comprehension • aptitude and personality

The study also found that the examiners may make judgements with different weightings of assessment criteria. Among the five assessment criteria, Presentation & Delivery (30%) and Fidelity & Completeness (56%) combined account for 86% of the 300 decisions made in the paired comparisons. Therefore, they can be regarded as the primary criteria that the examiners used. The two criteria fall nicely in the two core layers – *accurate rendition* and *adequate expression* – of Pöchhacker's (2001) model of the quality standards for interpreting.

The other criteria, such as the Interpreting Skills & Strategies and the Audience Point of View, may fit into the two outer layers of Pöchhacker's (2001) quality model: *equivalent effect*, and *successful communicative interaction*. However, these criteria are more difficult to operationalize in an interpreting examination due to the contextual restrictions in an artificial examination situation. For instance, there is usually no real audience in the examination room. This is probably one reason why most examiners relied more on the two primary criteria in the interpreting examinations.

The main criteria used by the examiners and their judgement results were cross-examined in order to answer the second question, i.e. to ascertain the relationship between the use of assessment criteria and the students' ranking patterns. The answer is yes and no. Yes, the students' ranking patterns resulted from the examiners' use of the assessment criteria, in many cases the examiners used the same criteria and made similar judgements. However, the answer is also no because it was found that the examiners' judgement approaches varied in terms of how they applied the assessment criteria, such as attaching different weightings to certain criteria properties. Therefore, some examiners might apply similar criteria but made contradictory judgements, or used different criteria but still picked the same winners. In some cases, all the criteria used and the judgements made differed between the examiners. These variations appeared in examiners of all backgrounds, but more evident in interpreter examiners. For example, the interpreter examiners paid more attention to the Interpreting Skills and Strategies, which is probably due to their interpreter background and professional habit of teaching interpreting.

When scrutinising the comments of the thirty examiners, some interesting assessment behaviours were also noted that may affect the examiners' judgement results. In the next section, we shall explore those assessment behaviours to clarify what the reasons are for the examiners' inconsistent judgement patterns and their use of assessment criteria.

4.3 Qualitative results – examiners' behaviours

A range of examiner behaviours have been noted in this study, from the observable external behaviour, such as the use of assessment tools, to the internal behaviour, which is less straightforward to observe as it is in the mind of the examiners. Due to the limited space of this chapter, we will only summarise the most salient behavioural aspects that affect the examiners' judgements in interpreting examination.

As far as the external behaviour is concerned, this study found that when assessing interpreting, interpreter examiners tend to depend more on their professional skills by listening and taking notes, whereas non-interpreter examiners tend to rely more on the speech script. Nevertheless, many examiners from both backgrounds found the speech script useful in checking the content accuracy of the student interpreters' interpretation. This is probably because the speech script can help lighten the memory and cognitive workload when assessing simultaneous interpreting.

The internal behaviours were inferred from the comments of the examiners when comparing the student interpreters. This study found that the examiners in general followed a similar approach in deciding the winners. Normally, the examiners would first look at the fidelity and completeness of the student interpreters' interpretations. When two students' performances were similar to each other and difficult to separate using the FC criterion, the examiners would then make the final decision by considering the way the students' interpretations were delivered. This process is described here as the FCD approach. The FCD approach may also be the main factor that maintains the consistency of most examiners' judgement results as a group because 86% of the decisions were made based on the two primary criteria as mentioned earlier.

Although the FCD approach is common to most examiners, variations in judgements do occur. It was found that some examiners can make different judgements when looking at the same interpreting performance, and that some could make inconsistent judgements even when they were based on the same assessment criteria. There are other factors at play here.

Three main types of examiners' internal behaviours were identified in this study, which may have an adverse influence on the reliability of the judgement in simultaneous interpreting examinations. They are examiners' *attention*, *bias*, and *professionally-referenced standards* or *professional habit*. These examiner behaviours, or factors, play an active role in the variations in examiners' judgements.

Assessing simultaneous interpreting imposes complex and high cognitive workloads on the examiners, monitoring two languages, making judgement and giving marks at the same time. Due to the limited attention span, therefore, examiners are likely to make the judgement by impression, i.e. holistic judgement, which affects the consistency of the judgement results. Examiner bias, such as personal preferences for the style of delivering the interpretation, is especially powerful in affecting an examiner's judgement. Some examiners show more tolerance towards a nervous but still faithful interpreter, while some others may react strongly to an interpreter whose delivery is jerky, or whose voice and expressions are perceived as annoying and irritating, which can only be a subjec-

tive viewpoint. All these will play a part in causing inconsistencies between the examiners' judgements (see Figure 4).

Interpreter examiners will also apply their professionally-referenced standards when assessing student interpreters. This is important because professional judgement is evidence of test validation for any performance assessment. When assessing student interpreters' interpretations, the examiners may refer to their personal experiences in the field and give different weightings to certain criteria according to the speech context for interpretation. Interpreter trainers tend to focus more on the problems of students' performances at different stages of training, whereas interpreter practitioners may give more consideration to the practical needs of the audience in the field. There are also some common norms. For example, when considering errors and omissions in the interpretation, generally speaking, both interpreter and non-interpreter examiners would rather that the student interpreters omit a message that is minor or not fully understood, than interpret it incorrectly and cause more confusion.

As Sawyer reported in his case study, however, "the [interpreter] jury members are a heterogeneous group in terms of professional experience as well as experience in teaching and testing" (2004, p. 184), it follows that the examiners' decision-making approach will inevitably be influenced by their different backgrounds and experiences. Although they may be following their professional norm to make judgements, those differences will play a subtle role in affecting the consistency of their judgements in the interpreting examination. Taking the interpreter examiners in this study as an example, overall they used similar assessment criteria and followed the FCD approach, but the between-examiner judgement patterns were evidently inconsistent. Based on the findings in this study as discussed above, the consistency and inconsistency of the examiners' judgement patterns appear mainly due to the examiners' various assessment behaviours and, perhaps, their different professional backgrounds.

5. A conceptual model of interpreting examinations

In order to gain a clearer perspective of how the above various elements work together, a basic conceptual model of interpreting examination (hereafter the basic IE model, see Figure 5) is proposed here to illustrate the relationships between the various elements in a typical interpreting examination.

In the IE model, an Interpreter Performance Scale is positioned at the apex of the Speaker-Interpreter-Audience triangle, or the assessment criteria triangle, with the two primary assessment criteria on the two slopes: Fidelity and Completeness (FC), and Presentation and Delivery (PD). The assessment criteria are

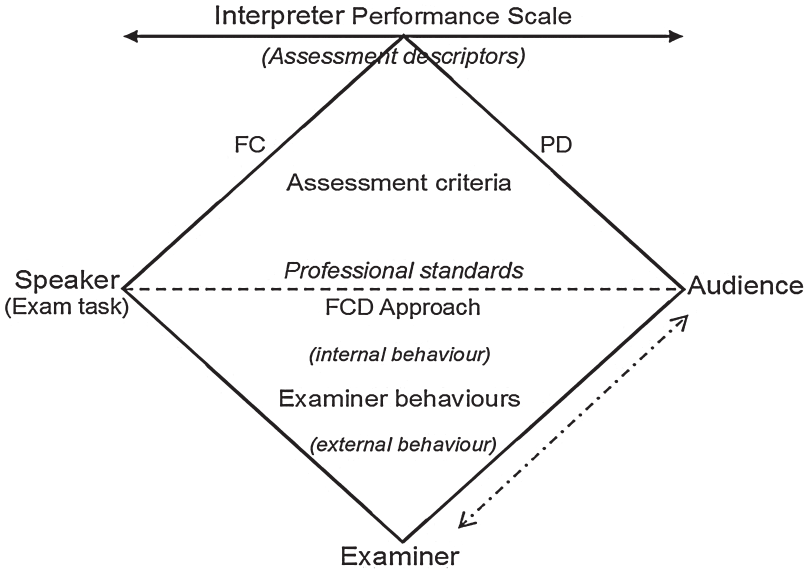
based on the Professional Standards (the base of the assessment criteria triangle) that interpreters follow when interpreting for the Speaker (Exam task) and the Audience. Interpreting examinations adopt these standards as assessment criteria for test validity reasons.

In the IE model, the factor of examiner behaviours is illustrated by a triangle of Speaker-Audience-Examiner, or the examiner behaviour triangle. The Examiner plays a dual role of an assessor and as a member of the audience. At the Speaker-Audience interface adjacent to the assessment criteria triangle is the FCD Approach. The Examiner usually follows this general approach to apply the assessment criteria, which is influenced by a range of external and internal examiner behaviours.

The IE model is represented as a dynamic and balanced system: the Interpreter Performance Scale, the assessment criteria triangle and the examiner behaviour triangle. The balance in the IE model is essentially maintained by a two-dimensional tension between the criteria triangle and the behaviour triangle, with the Examiner positioned at the bottom balancing everything on top in order to achieve a reliable test result.

In the assessment criteria dimension, the Examiner needs to find a balanced weighting of multiple criteria in order to make a judgement; whereas in

Figure 5. The basic conceptual model of interpreting examinations



the behaviour dimension, the Examiner's judgement alternates in a spectrum of assessment behaviours, ranging from external to internal. The external behaviour concerns the use of assessment tools that help check the fidelity and completeness of the interpretation, while the internal behaviour relates to the Examiner's ways of interpreting and receiving the messages based on their own personal preferences and professional experiences. Alternating between the two dimensions, the Examiner attempts to maintain a balance that is intricate and delicate in keeping the interpreting examinations reliable and valid.

One thing to note is that in the IE model the Interpreter Performance Scale is open-ended. The judgement on the quality of interpreting performance is contingent on the weightings of the criteria used by the examiners, such as the two primary criteria on the two upper slopes, and on assessment descriptors that are to be developed according to the purposes of the interpreting examinations. This open-ended conceptual design of the performance scale allows flexibility for developing assessment descriptors that are practical and operational for interpreting examinations with various situational themes and purposes.

6. Conclusion

Inconsistent judgements and, to borrow Sawyer's term, fuzziness in the use of assessment criteria by examiners, are two themes identified in this study. In this chapter, we have shown evident variations in the examiners' judgements (Figure 4), and identified five assessment criteria with various conceptual properties that the examiners used when making judgements on students' interpreting performances. This study also revealed some assessment behaviours which may help us gain a better understanding of how we assess students in the interpreting examinations.

As with all research, nonetheless, there are limitations to this study. Due to the complex issues underlying interpreting examinations, it would be difficult to investigate all the issues thoroughly and comprehensively in a single study. Therefore, this study has focused on the initial stage of exploring the intricate core issues of interpreting examinations.

In real life, for example, interpreting examination panels usually consist of multiple examiners. Therefore, one obvious limitation is that this study did not produce data on the interactions and influences among examiners that may be present like in a real-life examination panel. Therefore, the results of this study may not be directly generalised to real-life conditions where a number of examiners sit on the same examination panel.

Nevertheless, even in a multiple-examiner panel, examiners will first form their opinions alone after listening to the interpreting performances, and then proceed to discuss and agree the marks with the other examiners. It is more logical and practical to understand individual examiner's assessment behaviour before studying how a number of examiners interact with each other. Therefore, the findings of this study, based as it is on individual examiners' assessment behaviours, are useful in the sense that it investigated the initial stage of the examination panel before the examiners enter into discussions.

Based on the study findings, the proposed IE model can be viewed as a conceptual map to guide us through the fuzziness of the intricate relations between various elements in the interpreting examinations and the assessment behaviours of examiners. The proposed model, however, is by no means a final version, but should be regarded as a working model for continuing understanding of the interpreting examinations. The model can be used as a basic framework for further exploration and discussion on how these various elements may be operationalized in the interpreting examinations. There are still many knowledge gaps to be filled. For example, in the criteria dimension, works are still needed to find out how the test constructs and assessment criteria interact with one another; in the behaviour dimension, more studies are required to find out how the examiners' assessment behaviours can be considered in the test design, including the interactions between the examination tasks and the interpreter examinees. Studies like these can make contributions to producing useful test specifications for the interpreting examinations. With this knowledge, it is hoped that we can assess student interpreters in a way that is practical and reliable.

References

Alderson, J. C., Clapham, C., and Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Angelelli, C. V., and Jacobson, H. E. (2009). *Testing and Assessment in Translation and Interpreting Studies: A call for dialogue between research and practice*. Amsterdam and Philadelphia: John Benjamins.

Bachman, L. F., Lynch, B. K., and Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking *Language Testing, 12*, 238–252.

Bryman, A. (2004). *Social Research Methods* (2nd edition ed.). New York: Oxford University Press.

Caban, H. L. (2003). Rater Group Bias in the Speaking Assessment of Four L1 Japanese ESL Students. *Second Language Studies, 21*(2), 1–44.

Campbell, S., and Hale, S. (2003). Translation and Interpreting Assessment in the Context of Educational Measurement. In G. Anderman and M. Rogers (Eds.), *Translation Today: Trends and Perspectives* (pp. 205–224). Clevedon, UK: Multilingua Matters Ltd.

Eckes, T. (2005). Examining Rater Effects in TestDaF Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis. *Language Assessment Quarterly, 2*(3), 197–221.

Fulcher, G. (2003). *Testing Second Language Speaking*. Edinburgh Gate, UK: Pearson Education Ltd.

Hamp-Lyons, L. (Ed.). (1991). *Assessing Second Language Writing in Academic Contexts*. Norwood, N.J.: Ablex Publishing Corporation.

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

Kalina, S. (2005). Quality Assurance for Interpreting Processes. *Meta, 50*(2), 768–784.

Liu, M., Chang, C., and Wu, S. (2008). Interpretation Evaluation Practices: Comparison of Eleven Schools in Taiwan, China, Britain, and the USA. *Compilation and Translation Review, 1*(1), 1–42.

Lumley, T., and McNamara, T. F. (1993). Rater Characteristics and Rater Bias: Implications for Training, *Conference paper at the 15th Language Testing Research Colloquium*. Cambridge, UK.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

McNamara, T. F. (1996). *Measuring Second Language Performance*. London Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3 ed., pp. 13–104). American Council on Education, Washington: Macmillan.

Pöchhacker, F. (2001). Quality Assessment in Conference and Community Interpreting. *Meta, XLVI*(2), 411–425.

Pollitt, A., and Murray, N. L. (1996). What Raters Really Pay Attention to. In M. Milanovic and N. Saville (Eds.), *Performance, Cognition and Assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem* (pp. 74–91). Cambridge: Cambridge University Press.

Sawyer, D. B. (2004). *Fundamental Aspects of Interpreter Education: Curriculum and Assessment*. Amsterdam and Philadelphia: John Benjamins.

Thurstone, L. L. (1959). *The Measurement of Values*. Chicago: The University of Chicago Press.

Upshur, J. A., and Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing, 16*(1), 82–111.

Wu, S. (2010). Some reliability issues of simultaneous interpreting assessment within the educational context. In V. Pellatt, K. Griffiths and S. Wu (Eds.), *Teaching and Testing Interpreting and Translating* (pp. 331–354). Bern: Peter Lang.

Assessing Interpreter Aptitude in a Variety of Languages

Hanne Skaaden[1]
Oslo and Akershus University College of Applied Sciences (HIOA)

This chapter describes the admission test applied for enrollment in an undergraduate course on interpreting. The one year course (30 credits) prepares the students for the consecutive interpreting of institutional dialogues in the Norwegian public sector. Approximately 1100 candidates in more than 50 languages sat the oral admission test between 2007 and 2011. In the same period, an achievement test was administered for nearly 400 students in 49 languages upon completion of the course. The oral admission test is eliminatory and evaluates the candidate's performance in an interpreting-like task. A passing grade on the achievement test administered a year later, gives the student access to the National Register of practicing interpreters. In the qualitative analysis below, the admission test receives primary focus. Reporting on work in progress, the analysis describes the rationale behind the test design and the assessment scales implemented. The qualitative approach addresses validity and reliability issues, and describes measures taken to meet with comparability challenges across rater teams.

Key words: public sector interpreting, performance assessment.

1. Interpreting in the public sector setting – the need for testing

There is a registered need for interpreting in more than 100 languages in the Norwegian public sector, including the health and social welfare systems, the police and courts (IMDI, 2007). In the public sector setting, interpreting enables professionals and officials to inform, guide and hear parties in the case at hand, despite language barriers (Skaaden 2003, p. 74). Interpreting quality in this setting is accordingly of importance, not only for the minority speaker's legal safeguard, but also for the professional integrity of those in charge of the institutional discourse. Yet, interpreting quality is seldom addressed in research from this setting, Pöchhacker (2001, p. 420) notes. Grbić (2008, p. 239) points out that some authors who analyze interpreting in the public sector, even define the object of study outside the realm of interpreting to avoid dealing with quality issues. Some authors simply describe interpreting in the public sector as "non-professional" (Gonzáles, Vásquez, & Mikkelson, 1991, p. 29). In some studies from the public sector setting the discourse analyzed (cf., e.g., Meyer, 2001, p. 103) has little to do with the activity of interpreting as it is understood here, i.e., the activity of rendering in another language one person's speech (or utterances) for other listeners at the time the speech (or utterance) is made (cf., Ozolins, 1991, p. 39).

1 hanne.skaaden@hioa.no

Obviously, any assessment of interpreting quality and interpreter aptitude must confront a number of challenges due to the complexity of the skills involved in the activity of interpreting. In particular, challenges associated with the complexity of the phenomenon of bilingualism and with determining bilingual proficiency level (on bilingualism see e.g., Romaine, 1995; De Groot & Kroll, 1997; Wei 2005). These challenges are not particular for the assessment of interpreting quality in the public sector. Assessment challenges exist also in other settings where interpreting takes place, as pointed out by authors concerned with interpreting in international conferences (cf., e.g., Moser-Mercer, 1994; Kalina, 2002; Grbić, 2008). In the public sector the monitoring of interpreting quality and aptitude assessment are further complicated, due to factors such as professional secrecy and the array of languages needed (Skaaden, 1999a, 2003).

Seeking to meet with society's need for quality interpreting in a variety of languages, an assessment and accreditation system is under development in Norway. The system includes an accreditation test (established 1997, cf., Mortensen, 1998), a national register of practicing interpreters and their qualifications (since 2005, cf., Jahr et al., 2005), and an undergraduate training course (established 2007, cf., Skaaden 2008). The latter measure involves the admission and achievement tests addressed here. In the descriptive analysis, primary focus is placed on the admission test, with a brief description of the achievement test serving as backdrop. Through a descriptive analysis of the task design, the test administration and its rating system, the aim is to establish what factors may enhance or reduce the test's reliability under the existing circumstances? First, let us turn to a brief description of the rationale behind the test design.

2. The testing of bilingual skills

Any professional relies on a multiplicity of skills in carrying out the specific task assigned to his or her profession. In fact, the multiplicity of skills needed to perform it, is a prominent characteristic of a profession, the sociologist and philosopher Harald Grimen (2008, p. 72, 77) emphasizes in his discussion of the role of *know-how* and *knowledge* ("know-what") in professionalization. The interpreter profession does not differ from other professions in this respect. What distinguishes the *know-how* of the interpreter profession? It seems beyond dispute that a high level of bilingual proficiency is part of an interpreter's core competence. Accordingly, some sort of bilingual proficiency test is included in most aptitude tests for interpreter training. This practice corresponds with a widely held belief that a high level of bilingual proficiency is needed in order to perform the activity of interpreting professionally. In her

review of test methods applied for the testing of interpreter aptitude, Moser-Mercer (1994, p. 62) notes that written translation tests "often serve as a first hurdle before the candidate is admitted to the oral part of the aptitude test". She simultaneously points to skills needed in interpreting that cannot be tested through written translation, e.g., speed of comprehension, memory capacity, stress tolerance, voice and diction. In her discussion of the parameters and designs utilized in aptitude tests for interpreting courses, Moser-Mercer (ibid.) emphasizes that the nature and duration of the course should have impact on the choice of test design.

"Every test is an operationalisation of some beliefs about language, whether the constructor refers to an explicit model or merely relies upon 'intuition'", Alderson et al. (1995, p. 16) hold. On a general note, they (ibid. 187) stress that when it comes to the testing of linguistic skills "it is quite possible for a test to be reliable but invalid". While a multiple choice test may be highly reliable, performance on such a test reveals limited information about the candidate's ability to actually perform interpreting.

Performance tests (also known as 'authentic' tests), on the other hand, "use 'real-life' performance as criterion" (Davies et al., 2002, p. 144), and aim to assess "the ability of candidates to perform particular tasks, usually associated with job or study requirements" (loc cit.). In such tests "tasks are used to elicit language to reflect the kind of real world activities learners will be expected to perform", Wigglesworth (2008, p. 112) notes. She (ibid. 2008, p. 111) observes that over the past couple of decades "performance assessment has become increasingly important". In the same volume, Inbar-Lourie (2008, p. 286) ties this turn of development on the language assessment scene to the post-modern perspectives on knowledge as "an evolving, individually and contextually formed entity". Hence, we are witnessing an assessment culture where "the positivist traditional conceptions of validity and reliability" are being reconsidered (op cit. 287–288).[2]

A type of performance test frequently used as the first hurdle into studies of interpreting are written translation tests, as noted above. Such a test may reveal characteristics of the candidate's bilingual lexicon, grammars and writing capacity in the working languages. Written tests have limitations in testing for interpreter aptitude, however, as they do not reveal the candidate's pronunciation skills or how well the candidate's bilingual abilities hold up under the time pressure. These are interesting aspects in interpreting where the source

2 The development on the assessment arena reflects, I believe, a change of focus within general linguistics, where more attention is paid to the dependency of linguistic meaning on context and interactivity (Linell, 2009).

utterance is presented only once, giving the interpreter little room for analysis, before he must repeat the utterance in the other language. The extreme time-pressure and the fact that the interpreter cannot analyze the "text" as a whole, distinguishes interpreting from written translation, the translation scholar Otto Kade pointed out (Pöchhacker 2004, p. 10–11). The specific work conditions pointed to by Kade can be related to the theoretical considerations of Gile's *Effort Model* (, 1995; 2009). Given the challenging circumstances under which the interpreter's working memory performs, the *effort model* sees the activity of interpreting as relying on a type of mental energy or resources. It further postulates that if the resources required (RR) surpass the resources available (RA) so that RR > RA, then performance will deteriorate.

The one-year training course that our candidates wish to enroll in is not a language course, and those admitted should have a bilingual proficiency level that allows them to take part in oral exercises in both working languages from the start. The fact that successful completion of the training course allows the students into the National Register of Interpreters[3] makes an oral performance test a suitable design for our purpose. Accordingly, the admission test aims at eliminating candidates whose bilingualism is not yet strong enough to follow the course's learning activities. The need to test bilingual skills in an array of languages, presents a test design with challenges regarding task construction and rating consistency, however. The question addressed in the following is, thus: how does the current test design meet with the challenges presented by the current conditions?

3. The assessment of interpreter aptitude – the Norwegian experience

The training course for which the admission and achievement tests are implemented, prepares the students for interpreting in the Norwegian public sector. For pedagogical reasons the course focuses exclusively on the consecutive interpreting of institutional dialogues (Skaaden, 2008). In line with models of on-site interpreting as the pursuit of both rendering utterances in another language and coordinating verbal interaction (Wadensjö, 1998), the students during the course exercise interpreting strategies, and engage in learning activities aiming to increase their bilingual sensitivity. Here, focus is placed on the admission test's task design, test administration and rating system.

3 The register is publicly available at http://www.tolkeportalen.no (Downloaded 12-01-2012)

3.1 The admission test

The admission test screens the candidate's ability to render an utterance heard in one language in the other working language shortly after. Paraphrasing is usually associated with repetition of an utterance in the same language that it is presented to you, and Moser-Mercer (1994, p. 63) describes "oral *paraphrasing* of a message received orally" as a task "commonly used to test for native language competence." In our case the candidate must repeat the utterance in his other working language without much delay. Hence, the test with relative authenticity mimics the task of consecutive interpreting in dialogues. The task is performed in both language directions. A similar task design is employed by the *L'École Supérieure d'Interprètes et de Traducteurs* (ESIT) in their Épreuves d'admission, where the candidate must recapitulate an exposé heard in one language in the other working language shortly after, and the candidate's linguistic performances are evaluated according to his ability to analyze incoming information and express himself in both language directions.[4] In our case each sequence is shorter than in the ESIT test, but the tasks correspond through the basic activity of listening in one language and repeating in the other.

3.1.1 The task and the candidates

The task consists in the test examiner presenting the candidate with a short exposé, sequence by sequence. The candidate must immediately repeat each sequence in the target language. Subsequently, the same task type is repeated in the other language direction with a different exposé. The exposé informs on a topic relating to health, societal or legal issues, but in a non-specialist, general type of vocabulary. For instance, its topic may be 'on people's lifestyle and health', 'children's dental care', 'advice to victims of domestic violence'. The texts serving as exposé consist of 12 to 15 sequences that follow a logic string to create coherence. At the same time, each sequence has independent content so that the candidate's working memory has something to "grasp".

The sequences presented to the candidate are relatively short (1 to 3 sentences or 5 to 15 meaning units), e.g., "[Now I will share with you] [some information on] [what to do if] [you have been exposed to] [domestic violence]". Although the exposé is written down in order to secure comparability across candidates and languages, care is taken to make the sequences' syntax and word order "oral". The examiners are also instructed to "speak" the sequences,

4 A detailed description of this part of the ESIT Épreuves d'admission is found at http://paiement-esit.univ-paris3.fr/modalites/view/inter/page:4 (Downloaded 12-01-2012)

and not read them. This is of importance, as we know that even experienced conference interpreters have problems grasping text that is being read (Anderson 1994, p. 104).

To give the reader an impression of the task, excerpts from two different candidates' responses to the same exposé are included as excerpts (1) and (2) on the next page. Each example includes three out of the particular exposé's 14 sequences that were heard in Persian and were to be rendered by the candidate in Norwegian. The test conditions are identical in the two cases. Both candidates are adult migrants whose L2 is Norwegian. As we can see, their performances differ considerably under equal conditions. While the first candidate (1) succeeds in rendering the source speaker's utterances, the same task presents the candidate in (2) with severe problems, and key concepts, coherence and meaning are lost. We return to the excerpts and the scores in subsection 4.

All candidates are recruited through newspaper and internet ads. Candidates with previous interpreting experience and candidates without such experience are treated equally in the test. The general impression is that candidates of both backgrounds can handle the test conditions, and successful completion of the test does not depend on previous experience with interpreting. In fact, such experience is no guarantee for a successful result. For instance, the candidates in excerpts 1 and 2 both report to have previous experience as interpreters in the public sector.

Excerpt 1

Content heard in SL	Candidate 1's rendition in TL	Score
Several researchers have carried out a study on people's life style and health condition	Some researchers have carried out research on health and the life style of a group of people	6
The researchers focused on the relationship between smoking, alcohol use, physical activity and the intake of fruits and vegetables	The researchers concentrated on the subjects' smoking, alcohol use, physical activity and erm erm their intake of vegetables	5
The researchers found that even small changes in your life style could improve your health considerably	The researchers have reached the conclusion that erm physical activity and erm small changes of life style may help you towards better health	5

Excerpt 2

Content heard in SL	Candidate 2's rendition in TL	Score
Several researchers have carried out a study on people's life style and health condition	Some – or erm erm have done research on erm on erm – could you repeat (xxx) [source speaker repeats] Some researchers have researched how erm erm erm is it how it is to live healthy and erm …	3 (1)
The researchers focused on the relationship between smoking, alcohol use, physical activity and the intake of fruits and vegetables	Researchers have erm erm looked very much at how it has been to/ with smoking and erm erm to eat vegetables and erm – to drink alcohol and have physical activities	3 (1)
The researchers found that even small changes in your life style could improve your health considerably	The researchers came to [i.e. found] that that even small adjustments and changes may make up very big changes in how you erm erm live or …	3 (1)

Bilingualism is always a matter of degree (Romaine, 1995), and for bilingual speakers the concept *native language* is not a clear-cut category. Moreover, first language attrition may take its toll on the bilingual profile of long term migrants (Skaaden, 1999b). Hence, the equal testing of linguistic skills in both working languages is necessary. The current candidates' bilingual profiles display variation, and some characteristics should be emphasized. A vast majority (85–90 %) of the candidates that sit and pass the admission test are adult immigrants to Norway. Accordingly, Norwegian is their second or "B" language. As for the remaining 10–15 percent a sub-group is formed by native speakers of Norwegian who have acquired a second working language abroad. Candidates who are raised in Norway, but with their immigrant parents' native language as "home language" form another subgroup.[5] A unique position is held by students of the Saami languages as speakers of Indigenous languages of Norway. Their bilingual profile, with a life span development in a stable bilingual community, differs in type from bilingual speakers who acquired their second language after adolescence or in a migrant community with bilingual instability (cf., Romaine, 1995; Hyltenstam & Obler, 1989).

5 Candidates from this sub-group may have problems passing the test due to limited exposure to their language other than Norwegian. As one young girl testing for Albanian said upon completing the test: "I'm sorry, this Albanian is too adult for me!" Her Norwegian seemed flawless.

3.1.2 Test administration and rating

The test is administered face to face or by telephone. In the latter case the candidate performs from a public office (e.g., a police station or an embassy) that controls the test conditions. A test team of two, a Norwegian speaking test administrator and an ancillary bilingual examiner, administer the test for each language group. The Norwegian speaking examiner administers tests in co-operation with ancillary examiners in a number of languages. After a brief presentation of the two examiners, instructions are communicated to the candidate by the test administrator. The candidate is routinely informed that:

- *Before the test starts, I will explain to you how the test is carried out*
- *This is a listening test, so you are not to take notes*
- *You are to listen to a text/exposé in Norwegian, and repeat what you hear in [name of the target language, e.g. Persian] when I/the speaker pause(s)*
- *You will hear relatively short sequences at a time, and you are to repeat the sequence as soon as I/the speaker pause(s)*
- *Place your focus on rendering the content of the sequence – as fully as you can in the other language.*
- *You may ask for the sequence to be repeated once, if you have not understood or do not remember everything.*
- *Did you understand the task? Do you have any questions? … (if questions, they are answered before moving on)*
- *The test will be tape recorded. The subject of the exposé that you will hear in Norwegian is … (e.g. "advice to a victim of domestic violence")*
- *If you are ready, we now proceed with the test…*
  As soon as the candidate completes his/her rendition of the exposé from Norwegian into the target working language, the test proceeds in the other direction. The test administrator, accordingly, informs the candidate that:
- *We proceed immediately in the other language direction. You are now going to listen to a text/exposé in … [e.g. Persian] and repeat what you hear, sequence by sequence, in Norwegian. The topic of the next exposé is …. (e.g., "lifestyle and health condition"*)

While the Norwegian speaking examiner presents the first exposé in Norwegian, the bilingual ancillary examiner rates the candidate's renditions in the target language – sequence by sequence. In the second part, the examiner/rater roles change. The ancillary examiner is a native speaker of the language (other than Norwegian) being tested. The need to test in a large number of languages implies that "trained linguists" cannot be found in all language combinations sought for. Often, other practicing interpreters are also not adequate for this role, due to

"legal incapacity" in a tough and narrow market like the Norwegian. The solution in our case has been to build a network of bilinguals from different professions, and guide and instruct them in the task.

The candidate's rendition of each sequence is rated on-site according to the following analytic rating categories:

A. *Rendition in the other language*
i: A very good rendition with distinctions intact (6 points)
ii: Few/minor imprecisions in the rendition (5 points)
iii: Several or severe inaccuracies in the rendition (3 points)
iv: The meaning partly disappeared in the rendition/The candidate was unable to render everything (1 point)
v: The meaning was completely lost in the rendition/The candidate was unable to render the sequence (0 points)

The raters are instructed to evaluate the candidate's performance in each sequence individually and according to the qualitative category descriptions. Subsequently, the qualitative ratings are translated into a quantitative scale (6–0 points).

The raters have access to a key text during rating. They are, however, instructed to rate according to what the source speaker actually vocalizes, and not what the key text suggests. If discrepancies are suspected between the key and what the examiner actually said, the team must check whether "flaws" in the candidate's rendition are due to features of the source. Hence, if the source speaker skips or adds a meaning unit, the candidate's rendition in the target language must be rated accordingly. Second markings based on the recording are randomly carried out, and we return to this aspect in Section 4 below.

For an exposé with 14 sequences, the maximum score will be 84 points (6 x 14). The result achieved on the rendition task (A) is weighted with two other ratings: the raters must indicate their general impression of the candidate's *pronunciation* and *grammar* separately, according to the following scales:

B: Pronunciation
Native like/Very good pronunciation (100–85 %)
Clearly perceptible accent, but not difficult to understand (84–75 %)
Disturbing accent/accent that severely reduces understandability (74> %)

C: Grammar/phrasing
Native like/Very good grammar/phrasing (100–85 %)
Some mistakes, but not difficult to understand (84–75 %)
Disturbing mistakes/mistakes that severely reduce understandability (74> %)

The indication in percentage for the categories of B (pronunciation) and C (grammar/phrasing) is an impressionistic measure that enables the examiners to arrive at a total score for the performance. In the total score, the rendition task (A) weighs double (50 %) while the scores on pronunciation (B) and grammar or phrasing (C) weigh 25 % each. Accordingly, if a candidate scores 86 % on the rendition of sequences and has a native-like pronunciation and near native-like grammar (100 %+90 %), the candidate's total score will read 90,5 % in this language direction (i.e., 86+86+100+90/4). If the same candidate's total score in the other language direction reaches 80 %, the candidate will qualify for admission with a final score of 85,3 %.

Based on the scores arrived at during the test, a separate committee subsequently draws the candidates who are admitted to the course. In general, candidates with a total cut-score above 80 % in both language directions are admitted to the course. To be admitted, candidates should not have scores below 75 % on any of the parameters just described, however. Nearly 1100 candidates in more than 50 languages sat the oral admission test between 2007 and 2011. The pass rate for the admission test is approximately 40 %. In the same period, the achievement test was administered for nearly 400 students in 49 languages upon completion of the course. To indicate the framework of which the admission test is part, a brief outline of the achievement test follows.

3.2 The achievement test

The achievement test evaluates the candidate's performance in a role-played institutional dialogue upon completion of the one year course. The test evaluates the candidate's (a) ability to render linguistic and pragmatic distinctions in both language directions during the consecutive interpreting of an institutional dialogue, along with (b) the ability to coordinate interpreted discourse (e.g., the use of clarification strategies, pronoun choice, turn-taking). The candidate's performance is here evaluated according to a descriptive scale, in line with the standard grading system A-F. The grades describe the candidate's ability to render semantic and pragmatic distinctions in both language directions and to produce a natural flow in out-put compared to the source. Furthermore, the candidates' turn-taking, voice and diction, posture, mimics, and gestures are rated as to what extent communication is disturbed or not by the student's strategy choices.

The results from the achievement test indicate that a majority of the candidates admitted to the course are able to follow its learning activities. The course's completion rate exceeds 80 % for all five year classes. A majority of the students complete the course with passing grades, while a marginal 2.5 % fail (F). The

final grades place themselves on a curve that is slightly skewed in the positive direction, in that more than 75 % of the students pass with the grade C or better.

4. Issues of validity and reliability

A performance test has high *face validity* when it "appears to measure the knowledge or ability it claims to measure, as judged by the untrained observer", Davies et al (2002, p. 59, 144) stress. According both to candidates' and observers' responses the test seems to have strong face validity. The admission test can also be claimed to have *content validity*, as it mimics the real life task being tested for. Since the admission test is administered in both language directions, the test situation reveals the candidate's listening and oral production skills in both languages while under the pressure of the test situation.

The test performance partly depends on the candidate's ability to create enough distance to the form of the source utterance, to render its contents in the target language. Since the candidate is not allowed to take notes during the test, the candidate's memory capacity under the specific circumstances, moreover, has an impact on the test result. So may the candidate's affective reactions to the test situation as such (most candidates show some signs of nervousness). In sum, the same aspects play a part in the activity of interpreting proper thus serve as empirical underpinning for the performance test. Theoretical underpinnings for the design are found in Gile's (2009) Effort Model, predicting that an imbalance between the cognitive resources required and those available will cause performance to deteriorate. Since both empirical and theoretical considerations contribute to the test's validity, the choice of a performance design as the one described seems justifiable for the current purpose.

In terms of validity, Alderson et al. (1995) emphasize the importance of choosing a test design that fits the purpose. However, they also stress the interconnection between the validity and reliability continua:

> In principle, a test cannot be valid unless it is reliable. […] On the other hand, it is quite possible for a test to be reliable but invalid. […] Thus multiple-choice tests can be made highly reliable, […] yet some testers would argue that performance on a multiple-choice test is not a highly valid measure of one's ability to use language in real life (Alderson et al., 1995, p. 187)

Although the admission test may fare well on the validity continuum, the need for testing in numerous languages raises issues of reliability. First of all, the circumstances create the need for a number of rater teams. Test reliability against the backdrop sketched above moreover relate to (a) *task*, i.e., the nature of the exposé and how it is presented in both language directions, and (b) *the raters' perfor-*

mance, i.e., their understanding of own task, judging abilities, rating consistency, etc. These aspects are interrelated, but let us first look at the task construction and administration.

Task related threats to reliability can be linked both to the exposé's content and form. To assure comparability across languages, the exposé presented in the language other than Norwegian is produced on the basis of a Norwegian text of the same length and character as described above. The bilingual who adapts the exposé into the other working language is made aware of the importance to present the candidate with idiomatic expressions, coherent sequences and an "oral" syntax. Moreover, care is taken to choose texts that do not seem awkward in the language in which the exposé is presented. If a concept is unknown in the target culture, this is to be changed into a concept that occurs in the language's convention. Accordingly, the "transliteration" or "explanations" of Norwegian expressions must be avoided. The process is carried out under guidance from the regular staff. This routine was chosen after a pilot where the ancillary bilingual examiner was asked to provide a suitable text in his or her language. The pilot texts provided by the ancillary examiners proved too difficult for candidates to handle under the test conditions. To assure an equal level of difficulty in both language directions and across language groups, the current strategy was implemented.

The exposé is to reflect a "non-specialist" vocabulary. What is a "difficult concept" in terms of bilingual competence is hard to determine, however. For instance, in the preparation of the test for the Polish group, the Norwegian term *enmannsforetak* ('sole proprietorship', 'one-man firm') was considered a difficult term. As it turned out, none of the candidates found this particular term troublesome (Polish workers in Norway often found their own one-man firm). The example illustrates one type of difficulty associated with the assessment of bilingual proficiency: in bilinguals lexical knowledge is domain related, and depends on each speaker's previous experiences (Romaine 1995, p. 18). Interestingly, a term that did cause problems for several Polish candidates was the Norwegian concept *skulder* ('shoulder'), as the conceptualization of the body parts (*arm vs. shoulder*) differs in the two languages. This example serves to indicate that characteristics of the two conventions also play a part when bilinguals make an effort to keep their linguistic registers apart under the stressful conditions of a test situation.

The use of analytical scales for the evaluation process has both advantages and disadvantages. An advantage is that the raters easily grasp what they are expected to do. Yet, the approach leaves room for subjectivity. Double marking, with the second marking based on the sound track, shows that ratings are not always consistent. The performance quoted in excerpt 2 above, went through double marking as indicated by the scores in parentheses "3 (1)". While the first rater marked the three sequences each with 3 points (i.e., *several or severe inaccuracies),* the second rater

marked the same sequences with 1 point each (i.e., *the meaning partly disappeared /the candidate was unable to render everything).* It would seem that the second rater is closer to target in his judgment. The example illustrates a general tendency observed in double rating that the second rater, judging on the base of the recording, is stricter than the on-site rater during the test. A possible explanation for this effect is that the second rater listens to the recording several times, thus, detects more details. In the particular case (cf. excerpt 2, above), both raters concluded that the candidate's performance did not qualify for admission, although they both rated the candidate's performance relatively high on pronunciation and grammar. Simultaneously, the example serves to reveal that the difference between the two qualitative rating categories just quoted is rather subtle and could be improved.

Due to the multitude of languages and language varieties involved in the current project, routines were developed for instructing and guiding the ancillary examiners and raters. In addition to the instructions given before the test takes place, the cooperation with the regular staff and test administrator throughout the process plays an important part in reaching reliable procedures. The reliance on ancillary examiners is not particular for the testing of potential interpreters, but pertains to the testing of bilingual populations in general. For instance, Noland's (2009) study is concerned with the psychological testing of bilingual students either by a bilingual psychologist or via a bilingual ancillary examiner. She concludes that it is of yet unclear as to how the ancillary examiners affect the test conditions (ibid. p. 43). Noland's results are not directly applicable to the present test. However, her finding (ibid. p. 40) that effects may be task related, are interesting.

Evidently, in our case the test's reliability depends to some extent on the successful cooperation with the bilinguals who aid in task preparation and serve as examiners/raters. With the constant need for new languages, the routines must be instructive for raters who do this for the first time. The problem may to some extent be balanced by the supervision of the test administrator who has experience from numerous language groups and can guide new ancillaries in the process. Based on the experiences described in this chapter, in the further development of the test and its administration, measures to enhance reliability should be accentuated, in particular in the areas of bilingual task construction and inter-rater consistency.

5. Conclusion

Inevitably, a high level of bilingual proficiency is a prerequisite for an interpreter's successful performance. The activity itself is characterized by the performer's ability to understand and speak two different languages, as well as keeping them apart and willfully switching between them under severe stress and time

pressure. Due to the complexity of the skills involved in interpreting, as well as the complex nature of the phenomenon that makes up the interpreter's basic tool – language, any test that sets out to assess interpreting skills will be faced with validity and reliability issues. In general, a performance test like the one described here fares well on the validity continuum in that it mimics a real life activity. However, such test designs release multiple challenges on the reliability continuum. The aim of this chapter has been to identify some of these challenges, and suggest measures to meet with them within the current framework.

An incentive for the present research is the recognition that the road to professionalization for interpreters in the public sector setting goes through stated quality requirements, regardless of language combination. At the same time, the assessment of interpreting quality and interpreter aptitude depends on multiple, and in themselves complex, factors which are difficult to standardize and expensive to control.

References

Alderson J. C., Clapham, C., Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Anderson, L. (1994). Simultaneous interpretation: Contextual and translation aspects. In S. Lambert & B. Moser-Mercer (eds.) *Bridging the Gap. Empirical research in simultaneous interpretation.* Amsterdam: John Benjamins. 101–120.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., McNamara, T. (2002). *Dictionary of Language Testing. Studies in Language Testing 7.* Cambridge: Cambridge University Press.

De Groot, M. B., Kroll, J.F. (1997). *Tutorials in Bilingualism. Psycholinguistic Perspectives*. Mahwah, N.J.: Lawrence Erlbaum

Gile, D. ([1995] 2009). *Basic concepts and models for interpreter and translator training. Revised edition*. Amsterdam: John Benjamins.

González, R.D., Vásquez, V.F., Mikkelson, H. (1991). *Fundamentals of Court Interpretation. Theory, Policy and Practice*. Durham, N.C.: Carolina Academic Press.

Grbić, N. (2008). Constructing interpreting quality. In *Interpreting, 10, (2)*, 232–257.

Grimen, H. (2008). Profesjon og kunnskap. [Profession and knowledge]. In A. Molander & L. I. Terum (eds.) *Profesjonsstudier [The Study of Professions].* Oslo: Universitetsforlaget, 71–87.

Hyltenstam, K., Obler L.K. (1989). Bilingualism Across the Lifespan. Aspects of Acquisition, Maturity and Loss. Cambridge: Cambridge University Press.

IMDI (2007). *Tolkeformidling i offentlig sektor. Etterspørsel og tilbud. [The Mediation of Interpreters in the Public Sector. Demands and Supplies.]* Oslo: IMDI-rapport.

Inbar-Lourie, O. (2008). Language Assessment Culture. In E. Shohamy & N. H. Hornberger (eds.). *Language Testing and Assessment. Encyclopedia of Language and Education. Vol. 7.* New York: Springer Science, 285–301.

Jahr, K., Karterud, T., Rachlew, A., Skaaden, H., Stangvik, G. K., Tuv, E. (2005). *Rett til tolk. Tolking og oversettelse i norsk straffeprosess. [The Right to an Interpreter. Interpreting and Translation in Norwegian Criminal Procedure.].* Oslo: Justis- og politidepartementet.

Kalina, S. (2002). Quality in interpreting and its prerequisites. A framework for a comprehensive view. In G. Garzone & M. Viezzi (eds.) *Interpreting in the 21st Century. Challenges and opportunities.* Amsterdam: John Benjamins, 121–130.

Linell, P. (2009). *Rethinking Language, Mind and World Dialogically. Interactional and contextual theories of human sense-making.* Charlotte: Information Age Publishing Inc.

Meyer, B. (2001). How Untrained Interpreters Handle Medical Terms. In I. Mason (ed.) *Triadic Exchanges. Studies in Dialogue Interpreting.* Manchester: St Jerome, 87–107.

Mortensen, D. E. (1998). *Establishing a scheme for interpreter certification. The Norwegian Experience. The Norwegian Interpreter Certification Examination.* Oslo: University of Oslo, Department of Linguistics.

Moser-Mercer, B. (1994). Aptitude testing for conference interpreting: Why, when and how. In S. Lambert & B. Moser-Mercer (eds.) *Bridging the Gap. Empirical research in simultaneous interpretation.* Amsterdam: John Benjamins, 57–69.

Noland, R. M. (2009). When No Bilingual Examiner Is Available: Exploring the Use of Ancillary Examiners as a Viable Testing Solution. In *Journal of Psychoeducational Assessment 27(29)*, 29–45.

Ozolins, U. (1991). *Interpreting, translating and language policy.* Melbourne: National Languages Institute of Australia.

Pöchhacker, F. (2001). Quality Assessment in Conference and Community Interpreting. In *Meta, 46, (2)*, 410–425.

Pöchhacker, F. (2004). *Introducing interpreting studies*. London: Routledge.

Romaine, S. (1995). *Bilingualism.* 2nd Edition. Oxford. Blackwell.

Skaaden, H. (1999a). Lexical knowledge and interpreter aptitude. In *International Journal of Applied Linguistics*, *9, (1),* 77–99.

Skaaden, H. (1999b). *In Short Supply of Language. Signs of First Language Attrition in the Speech of Adult Migrants*. Oslo: Unipub forlag.
Skaaden, H. (2003). On the bilingual screening of interpreter applicants. In Á. Collados-Aís, M.M. Fernándes Sánchez & D. Gile (eds.) *La evaluación de la calidad en interpretación: investigación*. Granada: Comares, 73–85.
Skaaden, H. (2008). Interpreter students interacting in the cyberspace classroom. In C. Valero-Garcés (ed.) *Research and Practice in Public Service Interpreting and Translation. Challenges and alliances.* Alcalá: Obras Colectivas Humanidades 10, 169–182.
Wadensjö, C. (1998). *Interpreting as Interaction*. London: Longman.
Wei, L. (2005). *The Bilingualism Reader*. London: Routledge.
Wigglesworth, G. (2008). Task and Performance Based Assessment. In E. Shohamy &. N. H. Hornberger (eds.). *Language Testing and Assessment. Encyclopedia of Language and Education. Vol. 7.* New York: Springer Science, 111–123.

Unpacking Delivery Criteria in Interpreting Quality Assessment

Emilia Iglesias Fernández[1]
Universidad de Granada

This article uses findings from survey and experimental research on quality in simultaneous interpreting to show the complex mechanisms governing quality perception and assessment in interpreting. These studies are helping researchers unpack quality criteria in an attempt to establish more reliable assessment instruments in interpreting. The perception of interpreting quality seems to defy the atomistic notion of quality as expressed in the traditional lists of individual criteria. Findings have shown how the perception of one individual criterion crosses perceptual boundaries and impacts the assessment of other criteria, resulting in perceptual interrelations. This is particularly the case of delivery criteria: fluency, intonation, pleasant voice, native accent and diction. Additionally, in the process of interpreting assessment, the standard of delivery seems to be instrumental in the perception of accuracy, reliability and overall quality, in a process where the makeshift line that separates the assessment of message content from the assessment of delivery becomes blurred. Delivery criteria emerge as critical components in assessment, and yet, they are still not well understood, and remain largely intuitive. Studies in quality assessment are shedding some light on the perceptual nature of criteria and point to the need to revisit the existent delivery categories.

Key words: interpreting quality assessment, delivery criteria, perception, experimental studies, conceptualization studies.

1. Introduction

The aim of interpreting training is to help learners acquire interpreting skills so that they produce high quality performance. Assessment criteria play a major role in shaping quality performance by providing feedback in formative assessment, peer and self-assessment, and as a hallmark of professionalism in interpreter certification. It is therefore very important to review the quality assessment process and evaluation products to investigate whether assessment criteria measure the constructs they are intended to measure for the contexts and stakeholders involved, whether all raters involved understand the same thing by the criteria under assessment, and if the criteria used for assessment fully capture the phenomenon under study. Survey and experimental research in interpreting quality assessment are contributing to unravel some of the complex issues related to assessment in interpreting. Observation-based research on end-users of the interpreting services has focused on their abstract judgments about the relative weight

1 emigle@ugr.es

of a list of variables or "quality criteria" on interpreting quality (Kurz 1989, 1993; Vuorikoski, 1993; Kopczynski, 1994; Moser, 1995; Chiaro & Nochella, 2004). A second trend of quality studies has supplemented the traditional surveys of users which elicit their abstract priorities on quality with experimental studies which elicit users' assessment of genuine instances of interpretations. (Collados Aís, 1998; Pradas Macías, 2003; Collados Aís, Pradas Macías, Stevaux & García Becerra, 2007; Iglesias Fernández, 2007). Research on quality in interpreting has employed a taxonomy of parameters that has been in use since the end of the 1980s. These criteria have evolved from professional standards in interpreting and consequently are very intuitive. Many of these constructs are still not well understood, as they are very elusive and notoriously hard to define. And yet, they are frequently used in rubrics, evaluation grids and certification tests. These assessment criteria are used with a confidence which is hardly justified if we take into consideration that their accounts are largely intuitive.

The past years have seen an increased interest in the use of more reliable assessment methods and tools in interpreting evaluation (Angelelli, 2000; Campbell & Hale, 2003; Clifford, 2005; Angelelli & Jacobson, 2009). When performance assessment informs decisions about individuals such as test scores and interpreter certification, assessment must be held to a high standard of validity. In fact, the higher the stakes for assessment the more rigorous must be the evidence for the validity of the assessment. Performance assessment in interpreting is a very complex issue, as it involves the evaluation of a multifaceted task which is encompassed in a set of categories which in turn are analysed by more than one rater. In order to capture the interpreting phenomenon we must strive for reliable quality criteria.

2. The culture of interpreting assessment

Interpreting has evolved from a profession to an academic discipline. Thus, it is only natural that the professional approach to interpreting embodied in the Association Internationale des Interprètes de Conférence (AIIC), particularly its standards (1982) and school policy (Seleskovitch, 1968; Mackintosh, 1995) have greatly influenced the didactics and the evaluation of interpreting. AIIC's views on quality and assessment are based on several assumptions. One of them involves the procedures governing interpreting assessment. According to this, practicing interpreters are in a better position to gauge quality and assess performance (Seleskovitch, 1968; Mackintosh, 1995). This is a long standing claim in interpreting training and assessment. Currently material selection, test administration and test design are at the hands of trainers/practitioners. They have used AIIC's

professional standards (AIIC 1982), their experience as practitioners, input from colleague trainers and fellow practitioners as well as opinions from end-users as a source to establish criteria for assessment. It is also common practice amongst training institutions to invite practitioners as panel raters at final examinations in undergraduate and master programmes. Trainers play a major role in assessing learners' performance. However, interpreting trainers and practitioners involved in assessment tasks are not required to have knowledge of testing theory and testing methods. This approach to assessment in interpreting contrasts sharply with the culture of assessment found in second language acquisition, which enjoys a long tradition of empirical research in psychometrics which informs testing and evaluation (Clifford, 2005).

Another of AIIC's views on evaluation in interpreting pertains the components that capture the interpreting phenomenon. We are referring to the quality criteria against which candidate interpreters' performance is measured. An overview of the literature of the didactics of interpreting assessment shows that the majority of existent criteria stem from AIIC's professional standards (AIIC, 1982, Hartley et al. 2004). The "official" approach to assessment has also influenced the methodology of survey research on quality expectations and assessment in interpreting. Evidence of this can be found in the 1986 pioneering study on quality expectations conducted by Bühler which has greatly influenced subsequent studies on quality. In her work on quality views by practicing interpreters, Bühler (1986) nurtured on AIIC's standards (AIIC, 1982) and established nine quality criteria which she divided in two dimensions. One dimension relates to the linguistic (semantic) components of an interpretation and is made up of: *logical cohesion, accuracy, completeness, correct use of grammar, correct use of terminology,* and *appropriate style.* The second dimension depicts the extralinguistic (pragmatic) components, such as *native accent, pleasant voice,* and *fluency of delivery* (ibid: 231–232). Later survey works which shifted the focus from interpreters to end-users saw considerable modifications of the criteria denominations and the labels which couched the quality criteria but despite these proliferation of denominations a core of basic constructs has remained stable. The resulting taxonomy has remained more or less stable across studies of users' priorities (Kurz, 1989, 1993; Vuorikoski, 1993; Mack & Cattaruzza, 1995; Kopczynski, 1994; Moser, 1995; Chiaro & Nocella, 2004), as well as works concerned with users' assessment (Collados Aís, 1998; Pradas Macías, 2003; Garzone, 2003; Collados Aís et al., 2007). These criteria have also nurtured the items in surveys on customer satisfaction (SCIC, 2007, 2010). Additionally, they are frequently found in the various assessment instruments at different levels of training and professionalization: from feedback sheets (Schjoldager, 1996; Cenkova, 1999; Riccardi, 2002; Martin & Abril, 2003; Hartley, Mason, Peng & Perez, 2004) in the early stages of learn-

ing to assessment grids proper (EMCI, 2001) and interpreter monitoring systems (SCIC's Système d'Enregistrement de Rapports sur les interprètes Freelance).

Training institutions across the world adopted this set of criteria. Most CIUTI members (International Permanent Conference of University Institutes of Translators and Interpreters) and the European Master in Conference Interpreting (EMCI, 2001) use them. So do professional organisations such as AIIC's Admission Committee and the European Commission's Directorate General for Interpretation (SCIC). SCIC has recently launched an interpreter monitoring mechanism, SERIF (Système d'Enregistrement de Rapports sur les interprètes Freelance), which employs some of these assessment criteria to evaluate the competency of non-permanent interpreters working at the European Commission. SCIC also employs this set of criteria in its survey questionnaires used to measure customer satisfaction amongst European Members of Parliament (SCIC, 2007, 2010).

The catalogue of assessment criteria varies from training institution to training institution, and from country to country. Most evaluation grids in interpreter training (Schjoldager, 1996; Cenkova, 1999; EMCI, 2001; Riccardi, 2002; Martin & Abril, 2003; Hartley et al., 2004) cover quality along three axes, namely: message accuracy, command of the target language and effective delivery. Some, however, focus exclusively on linguistic and semantic features with no reference to extralinguistic factors (Falbo, 1998; The Languages National Training Organisation's National Standards LNTO, 2001) or apply some delivery criteria but are not exhaustive (EMCI, 2001). The majority exclude the interactional, pragmatic dimensions related to the situated context, as well as the affective factors so fundamental in public service interpreting. Consequently, there seems to be no consensus in the training literature on the number of criteria which would help to capture the interpreting phenomenon, nor is there agreement about the denominations used to couch the assessment constructs

Another of the central issues in valid and reliable assessment procedures lies in the definition of the categories for analysis, that is, determining what each of the categories that represent the construct measures. Categories should be clearly defined, and its subcomponents clearly spelled out, so that no ambiguity emerges. However, an overview of the existent assessment criteria in interpreting reveals a great degree of ambiguity, with definitions falling within two major categories.

While it stands to reason that different interpreting contexts and situations would call for different components at varying degrees, a core of basic assessment criteria seems to be at the root of all instances of interpreting, and these categories should be unambiguous. And yet, there is no one definition of *accuracy*, *fluency* or *pleasant voice*, to name just a few.

3. Criteria in interpreting quality assessment: shortcomings

Advancement in the unraveling of the assessment constructs in interpreting is slow, as spoken speech is elusive and defies measurement, but identifying shortcomings can contribute to elucidating the constructs. This section is devoted to presenting the reader with some of the shortcomings observed with regard to criteria in interpreting quality assessment. The shortcomings observed can be categorized according to the following factors.

One of the problems affecting quality criteria is that of the profusion of denominations; as various labels are used for one and the same criterion. Take for instance the case of the concept of equivalence (be it semantic, linguistic and pragmatic), which is usually represented with three different denominations, if not more. Labels such as *fidelity*, *sense consistency with the original* and *accuracy* coexist in the literature. The same applies to *pleasant voice* which receives four different denominations: *pleasant voice*, *voice*, *voice quality* and *voice qualities*. As to *fluency*, it is referred to as *fluency*, *rhythm* and *tempo*. Researchers use different labels when referring to the same criterion, and although some would claim that these denominations are synonymous, in fact they have been found to lead to different nuances in interpretations (Iglesias Fernández, 2006). This is particularly the case when the criteria are translated to other languages where the terms carry slight variations in meaning.

Another limitation affecting the existing criteria is that of the nature of their definitions. There are two broad categories of definitions for assessment components. Criteria are defined either according to their common-sense meaning or for their more technical sense. Let us take the criteria cited above. There are two possible definitions for *fluency*. One meaning is close to general proficiency in language, and the other is a more specialized sense, related to the temporal, suprasegmental features of speech, such as speech rate, uninterrupted runs of speech, number and duration of pauses (filled or unfilled), etc. As to *pleasant voice* or *voice quality*, as this parameter is usually referred to, its general meaning can elicit judgments for the voice pitch, intonation and voice volume whereas its technical meaning exclusively refers to features of the pitch (pitch range) and to voice timbre (lax or tense voice quality, creaky voice quality, breathy voice quality, etc.). The following question emerges: are respondent's judgments for a given criterion related to its common-sense meaning or are they to be taken for its more specialized sense? The profusion of denominations, coupled with the coexistence of general versus technical definitions for the criteria leads to varying interpretations and precludes the sharing and comparison between findings.

A serious concern arises from the use by researchers of one denomination to refer to different concepts. Bühler (1986), Kurz (1989, 1993) and Garzone (2003), for instance, use the denomination *pleasant voice* to elicit judgments for voice pitch, volume and intonation. The category intonation is not employed in their studies. Contrarily, Collados Aís (1998), Pradas Macías (2003, 2007) and Collados Aís et al. (2007) employ two criteria to encompass both the timbric and the prosodic features of the vocal flow. Thus the parameter *pleasant voice* is used to refer to voice pitch (fundamental frequency), its pitch range and its intensity (volume), whereas the criterion *intonation* is employed to depict the temporal characteristics of the pitch contour and pitch direction, that is, the voice melody. A very different stance is taken by Moser (1995), Russo (2005) and Kopczynksi (1994) who use a different denomination, namely, *voice quality*, to refer to the interpreter's pleasant voice. The label *voice quality* presents us with an additional problem: are these researchers using voice quality in its common-sense meaning of features of the voice or are they referring to its technical sense which alludes to features of the timbre? Do all respondents understand the same thing under this denomination? The denominations *voice quality*/*voice qualities* are compounded by perceptual limitations. Phonetic studies of the perception of the quality of the voice have shown that voice quality is an extremely difficult feature to grasp and to identify by naïve listeners, proving to be a very elusive feature even for experts (Biemans, 2000: 167). If these denominations do not allow for category identification, as the timbre features of the voice defy identification, it follows that respondents' answers for the *voice quality* criterion could be unreliable.

Clearly, assessment criteria in interpreting lack conceptual precision. Very often criteria which are listed as separate categories in some studies are presented grouped together under a general heading in other works. An example of this can be found in denominations such as *pleasant speech rhythm* (Vuorikoski, 1993), *monotonous intonation, monotonous tempo* (Kopczynski, 1994: 92), and *pleasant speech rhythm* (Mack & Cattaruzza, 1995) which group features of both intonation and fluency together.

This is further complicated by the use of these criteria to refer to native speakers as well as interpreters. Since *interpretese* has been found to be a particular feature of interpreting production, with a distinct intonation (Shlesinger, 1997) and fluency pattern (Pradas Macías, 2006). Should interpreters' fluency and intonation be assessed in terms of their resemblance to native source speakers or should their output be assessed in its own right?

4. Unpacking delivery criteria in simultaneous interpreting: the role of experimental research

Advancement in the unpacking of quality criteria in interpreting assessment has been made by the research group ECIS (Simultaneous Interpreting Quality Assessment) (Collados Aís 1998; Collados Aís et al. 2007; Iglesias Fernández, 2010; Collados Aís, Iglesias Fernández, Pradas Macías & Stevaux 2011; ECIS, 2012). ECIS has been actively involved in experimental work analysing the factors impinging on interpreting quality assessment. This research group has been particularly interested in the non-verbal dimension in interpreting. ECIS scholars were intrigued by the seemingly little importance that users attached to the non-verbal features in their quality preferences (Kurz, 1989, 1993, Kopzcynski, 1994, Moser, 1995). They noted that observational work on users' expectations fell short off the mark since it equated preferences with abstract, decontextualized opinions about interpreting. Furthermore, it did not take on board actual instances of assessment of interpretations. They introduced a methodological twist that moved empirical research on quality in interpreting a step further. The traditional survey research on quality expectations was supplemented with experimental assessment studies amongst users. Users were exposed to accurate interpretations that were rendered each with a manipulated poor delivery (a slight monotonous intonation, an unpleasant voice, a broad foreign accent, a disfluent speech, an unclear diction) (Collados Aís, 1998; 2007, Pradas Macías, 2003, 2007; Iglesias Fernández, 2007; Blasco Mayor & García Becerra, 2007). The same users who had regarded the interpreter's paralinguistic features (voice, intonation, accent, fluency, etc.) as not bearing much importance on quality in the expectations survey were nonetheless heavily influenced by the degraded delivery in the experimental study.

While cumulative evidence had shown abstract quality expectations impinging more on linguistic and semantic components, and very little on nonverbal components (Bühler, 1986; Kurz, 1989; Kopczynski, 1994; Moser, 1995), findings from users' actual assessment revealed that their satisfactory judgments of interpreting quality impinged largely on an effective and confident delivery (Collados Aís, 1998; Pradas Macías, 2003; Collados Aís et al., 2007; Iglesias Fernández, 2007). In fact, in simultaneous interpretation, weaknesses in delivery, particularly intonation, fluency, pleasant voice, and diction have extensive negative effects on delegate's overall assessment of the quality of the interpreter's output, and most strikingly, on the impression of accuracy of information (Collados Aís, 1998, 2007; Pradas Macías, 2003; Iglesias Fernández, 2007).

Collados Aís (1998, 2007) pioneered this methodological innovation when she studied the relative weight of two quality variables: the interpreter's monotonous

intonation and the interpreter's *accurate* rendition on the perception and assessment of an instance of simultaneous interpretation. Each variable was degraded in a controlled environment in two assessment studies. In the first experiment, users perceived and identified the interpreters' monotonous delivery, which resulted in very poor ratings for the *intonation* criterion. Surprisingly, the degraded intonation influenced users' low ratings for the *accuracy* of the message, a variable that had not been manipulated. In the second experiment, users identified the interpreters' vivid *intonation*, and rated it very highly, and thought the interpretation's content was accurate despite it not being consistent with the original message. Additionally, she noted that users' poor judgments for the interpreter's monotonous *intonation* not only reverberated in the perception of content *accuracy*, but also in their judgments for another temporal, paralinguistic variable, namely the criterion *pleasant voice*. Against this backdrop, the author concluded that informants' assessment of the quality of an interpreter's linguistic output and its sense consistency with the original can be very unreliable (Collados Aís, 1998; Gile, 1995), and that a lively melody in the interpreting ouput is instrumental in producing the impression of accuracy. Furthermore, she revealed that users had equated *intonation* with *pleasant voice* in their perception, and that it seemed to be extremely difficult to distinguish between the two in the perception process. A similar phenomenon was observed by another member of the ECIS group, Pradas Macías (2003, 2004) with respect to the criterion *fluency*. Pradas Macías noted the strong impact of a disfluent interpretation on users' perception of and judgments for *logical cohesion* and *intonation*. The critical effect of the interpreter's choppy delivery on the judgment for poor *accuracy* was also revealed. The disfluent interpretation was considered monotonous, lacking in logical cohesion and less consistent with the original message. These findings led the authors to arrive to two conclusions. One refers to the crucial and determining effect of delivery criteria on users' assessment of accuracy and overall quality. Secondly, they point to the complex mechanisms governing quality perception and its assessment, as perceptual interrelations amongst criteria occur in the assessment process.

These studies were replicated for 11 quality parameters on 197 end-users revealing a very similar outcome (Collados Aís et al., 2007). In the study on the relative importance of the interpreter's *unpleasant voice* on the assessment of quality, Iglesias Fernández (2007) noted how this delivery feature spilled on the perception of *intonation*, *fluency* and *diction*, as well as on the perception of *accuracy* and *overall quality*. Users' judgments for *intonation*, *fluency* and *diction* were much poorer when the interpreter's voice was unpleasant (Iglesias Fernández 2007). Blasco Mayor and García Becerra (2007) studied the role of a interpreter's degraded pronunciation (*diction*) and revealed that the interpreter's diction was equated with *fluency* and *pleasant voice*, after noting how users' judg-

ments for these two criteria were poorer in the face of the interpreter's unclear *diction*.

All these studies inform us that the abstract principle where quality results from a linear combination of information fidelity and delivery does not seem to hold. It is a proven fact that users have difficulties gauging information fidelity (Collados Aís, 1998, Gile, 1995) and interpreting quality (Collados Aís, 1998), and, deprived of access to the original, they are not necessarily sensitive to the content criteria. Users' satisfactory judgments of quality do not always correlate with the fidelity, linguistic acceptability or terminological accuracy of the output. Intuitive observation of interpreting practice had led researchers to notice how very often favourable impressions of accuracy were in fact informed by a pleasant and confident delivery despite a lack of sense consistency with the original (Namy, 1978: 26; Gile, 1991: 198). Conversely, the impression of message fidelity could be compromised when the output was delivered with a degraded paralinguistic quality (Gile, 2003: 112).

In fact, findings from ECIS have shown that raters do not seem to be in a position to assess each quality component separately as the perception of the quality of the message content is largely informed by clusters of superimposed delivery features (Iglesias Fernández 2010). This finding begs the question whether the traditional taxonomy of quality criteria in interpreting assessment is reliable and whether a study of the most conspicuous delivery clusters should be conducted. . The experimental studies aforementioned have shown that the perception of interpreting quality defies the atomistic perspective which is reflected in traditional taxonomies of criteria which present them as separate components. Quality categories do not seem to be processed separately, but holistically in clusters, a process where users' assessment for one criteria is likely to cross perceptual boundaries with other related criteria, particularly when the criteria share a common denominator. This has been shown to be the case of *fluency*, *intonation*, *pleasant voice* and *diction*, which are presented to the rater as separate categories. They all share the same medium, i.e. they are all temporal, suprasegmental variables of the interpreter's voice. The crossing of perceptual borders has also been observed between non-verbal features and the linguistic content.

The perceptual deficits and overlappings shown in these experimental works could explain the poor results of delivery criteria in users' quality preferences. Criteria that had repeatedly occupied the bottom positions in users' expectations such as *pleasant voice*, *clear diction* and *native accent* displayed a very high standard deviation in the studies by Collados Aís et al. (2007). A high standard deviation evidences a very heterogeneous response pattern, in other words, a high inter-rater variability. An overview of the response patterns of 190 users' expectations of quality revealed more inter-rater agreement for content criteria (very

low deviation): *sense consistency with original*, *completeness of information*, *appropriate terminology*, and less agreement (very high deviation) for form criteria of which *pleasant voice*, *diction* and *native accent* were affected by the highest deviation.

5. Unpacking delivery criteria: the role of conceptualisation studies

The presumable lack of agreement (very high standard deviation) in respondents' assessment for delivery criteria prompted two studies which aimed at eliciting users' definitions and mental representations for *pleasant voice* and *diction* amongst other parameters (Pérez-Luzardo, Iglesias Fernández, Ivars & Blasco Mayor, 2005; Iglesias Fernández, 2006). These categories were chosen amongst the criteria that seemed to give rise to perceptual overlappings in the experimental studies of quality assessment (Iglesias Fernández, 2007; Blasco Mayor & García Becerra, 2007). The ultimate goal was to probe whether the low ratings for these criteria in the expectations surveys were perhaps not the result of a methodological shortcoming, namely unsuitable criteria denominations or their co-occurrence with criteria which shared similar perceptual features. Two groups of respondents were asked to define what they each understood by either *pleasant voice* or *clear diction* (Pérez-Luzardo et al., 2005; Iglesias Fernández, 2006). The findings from these conceptual studies evidenced that some of the constructs being measured by these assessment categories tapped a different mental representation, that is, they were not actually measuring what they were supposed to. In the case of *pleasant voice*, respondents' definitions contained more allusions to the prosodic features of the voice, namely pitch contour and direction and tempo. No mention was made of its timbre and voice quality features (lax voice, tense voice, creaky voice, etc.). Users' definitions of *pleasant voice* were actually referring to the dynamic features of the voice, that is, *intonation* and *fluency*, and not so much to voice quality, intensity or pitch range, that it *pleasant voice*. This could explain why the original version of extralinguistic parameters by Bühler (1986) excluded the criterion *intonation* and only employed *pleasant voice* as the latter seems to subsume the former in user's mental representations. With regard to *clear diction*, participants in the conceptual mapping studies attached to *diction* properties more akin to *pleasant voice* (Pérez-Luzardo et al., 2005).

Similar phenomena had also been observed by Hartley et al. (2004) in a peer and self-assessment study involving interpreting trainees, trainers and end-users. The study aimed at eliciting criteria for assessment. In a pilot trial, trainees' prior awareness of existing criteria was elicited. The authors report misunderstandings

around the criteria commonly found in assessment grids. Thirty seven trainees, twenty two novices and fifteen advanced learners had difficulty defining delivery. As in the study by Pradas Macías (2003, 2004), the raters in Hartley et al. (2004) associated *fluency* with *cohesion* and with *coherence*. The pilot study on trainers and end-users revealed a similar pattern. *Accuracy* was often equated with *coherence*. Similar misunderstandings have been reported in the literature. Shlesinger (1997) mentions the confusion raised among participants in the survey study by Bühler (1986) around the category *pleasant voice*, particularly with the epithet "pleasant". One group of respondents in Bühler's study conjured up a sense of excellent delivery when they thought of "pleasant" voice whereas the other group of interpreters identified "pleasant" with average delivery (ibid: 197).

6. Conclusion

Assessment criteria play a major role in shaping quality performance in the different stages of training. They also provide a yardstick against which to measure professionalism in interpreter certification. Additionally, they are key instruments in eliciting client and end-user satisfaction judgements. In interpreting, evaluation procedures and products are still largely intuitive. They have been influenced by a certain culture of quality instilled by the profession according to which test design and assessment are better served if they are placed at the hands of trainers/practitioners. As to assessment criteria, they are largely a by-product of professional standards and have not been held to high standards of reliability. In fact, criteria have been found to be ambiguous. Constructs related to features of the delivery such as *fluency*, *pleasant voice*, *intonation* and *diction* are depicted by different denominations, leading to different interpretations. Thus, *rhythm* and *tempo* coexist with *fluency*. *Voice quality* and *voice qualities* are used alongside *pleasant voice*, and are found to refer to *intonation* too. As for *diction*, the labels *clear voice* and *pronunciation* are used. Conversely, different constructs have been found to be labelled under the same denomination. *Intonation*, for instance, is presented as a separate category in some assessment grids and feedback sheets, but is also found grouped under *pleasant voice* or *fluency*. Shortcomings also affect criteria definitions, as two broad categories of definitions seem to coexist; one category is closer to common-sense notions, the other is more technically oriented. *Voice quality* is a case in point. In its general meaning, *voice quality* refers to features of the pitch. This explains why the category *intonation* is also found couching this construct. In its technical sense, however, *voice quality* exclusively refers to features of the timbre.

All this conceptual shortcomings beg the question whether raters' responses to delivery-related criteria are uniform or whether judgments relate either to the general meaning or the technical sense. This situation leads to great rater disagreement which in turn limits the reliability of assessment products and evaluation result in interpreting.

Advancement in the refinement of assessment criteria is slow and is hindered by the intricate and elusive nature of the spoken speech in interpreting. Delivery-related criteria are particularly hard to define, and very difficult to distinguish as they all share a common medium, the voice. And yet the paralinguistic dimension of interpreting has been shown to be critical for favourable assessment of message accuracy and overall quality. This is especially true if we bear in mind that user's attention frays, and that users are poor judges of accuracy. It follows that delivery-related criteria should be unpacked. The study of these assessment components should move a step further, away from unidimensional stances and moving towards a more multidimensional approach. The overly simplified view of language and speech which permeates the culture of assessment in interpreting falls short of capturing the interpreting phenomenon.

One source of advancement in the unpacking of delivery criteria in interpreting assessment is found in studies concerned with exploring the relationship between the features of the interpreter's output and their perception by human agents. This line of research known as product oriented research in quality assessment is contributing to challenging existing assumptions on the assessment process and its products. Experimental studies have contributed to unravelling some of the intricate psychological mechanisms that govern interpreter perception and interpreter evaluation. Findings have evidenced that users do not seem to assess delivery criteria separately, but that *fluency*, *intonation*, *pleasant voice* and *diction* are perceived in clusters of features sharing a common denominator, the temporal dimension of the interpreter's voice. Additionally, these studies have revealed perceptual overlappings between presentation-related criteria. This crossing of perceptual boundaries does not only affect paralinguistic parameters, but has also been seen to interrelate formal features of the interpretation with the linguistic content, particularly with *accuracy*.

Conceptual mapping studies have revealed that the assessment category *pleasant voice* does not tap the mental representations which are intended to elicit, that is, pitch range, voice intensity and timbre. Instead *pleasant voice* conjures up the prosodic features related to *intonation* and *fluency*, that is melody (pitch contour and pitch direction) and tempo (speech disfluencies, speech rate). In much the same way, *diction*, evokes representations more related to the features of the *voice* and less to articulation. In light of these findings, caution should be exercised when using the *pleasant voice* criterion in interpreting assessment. *Pleasant voice* should not be

used alongside with *intonation*, since the former seems to subsume the latter. Similarly, caution should be advised when using *diction* alongside *pleasant voice*, as the latter seems to conjure up mental representations for the interpreter's articulation. These findings warrant further research which allows for larger samples of respondents to probe raters' mental representations for individual assessment categories.

This contribution has aimed to raise awareness of some methodological shortcomings and difficulties revolving around the existent delivery criteria in interpreting quality assessment. Refinement of presentation-related categories for assessment can help researchers investigating assessment in other environments, such as interpreter training and the certification of professional interpreters to make informed decisions about the use of assessment criteria.

References

AIIC (1982). *Practical guide for professional interpreters*. Geneva: AIIC.

AIIC Training Committee (2006). *Conference interpreting training programmes: Best practice*.

Angelelli, C. (2000). Interpreting as a communicative event: A look through Hyme's lenses. *Meta, 45*(4): 580–592.

Angelelli, C. and Jacobson, H. E. (2009). *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice*. Amsterdam/Philadelphia: John Benjamins.

Biemans. M. (2000). *Gender variations in voice quality*. Doctoral Dissertation. Katholieke Universiteit Nijmegen. The Netherlands, Utrecht.

Blasco Mayor, M. J. & García Becerra, O. (2007). La incidencia del parámetro dicción. In A. Collados, E. M. Pradas, E. Stévaux & O. García (Eds.), *La evaluación de la calidad en interpretación simultánea: parámetros de incidencia* (pp. 175–194). Granada: Comares.

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

Campbell, S. & Hale, S. (2003). Translation and interpreting assessment in the context of educational measurement. In G. Anderman & M. Rogers (Eds.), *Translation trends and perspectives*. Multilingual Matters: Clevendon.

Cenkova, I. (1999). Peut-on quantifier la qualité d'une prestation des étudiants en interprétation? Folia Translatologica, 6, 33–36.

Chiaro, D. & Nocella, G. (2004). Interpreters' perception of linguistic and non-linguistic factors affecting quality: A survey through the World Wide Web. *Meta, 49*(2), 278–293.

Clifford, A. (2005). Putting the exam to the test: Psychometric validation and interpreter certification. *Interpreting, 7*(1): 97:131.

Collados Aís, Á. (1998). *La evaluación de la calidad en interpretación simultánea. La importancia de la comunicación no verbal*. Granada: Comares.

Collados Aís, Á., Pradas Macías, E. M., Stévaux, E. & García Becerra, O. (Eds.) (2007). *La evaluación de la calidad en interpretación simultánea: parámetros de incidencia*. Granada: Comares.

Collados Aís, Á., Iglesias Fernández, E., Pradas Macías, E. M. & Stévaux, E. (Eds.) (2011). *Qualitätsparameter beim simultandolmetschen: Interdisziplinäre perspektiven.* Tübingen: Narr Verlag.

ECIS (2012). Quality in interpreting. In C. A. Chapelle. *The Encyclopedia of Applied Linguistics.* Oxford*:* Wiley-Blackwell.

EMCI (2001). European Masters in Conference Interpreting. Retrieved from http://www.emcinterpreting.org/assessment.php

Falbo, C. (1998). Analyse des erreurs en interprétation simultanée. *The Interpreters' Newsletter, 8*, 106–120.

Garzone, G. (2003). Reliability of quality criteria evaluation in survey research. In A. Collados Aís, M. M. Fernández Sánchez & D. Gile (Eds.), *La evaluación de la calidad en interpretación: docencia y profesión* (pp. 23–30). Granada: Comares.

Gile, D. (1991). A communicative-oriented analysis of quality in nonliterary translation and interpretation. In M. L. Larson (Ed.), *Translation: Theory and practice. Tension and interdependence* (pp. 188–200). Binghamton, N.Y.: SUNY.

Gile, D. (1995). *Basic concepts and models for interpreter and translator training.* Amsterdam & Philadelphia: John Benjamins.

Gile, D. (2003). Quality assessment in conference interpreting: Methodological issues. In A. Collados Aís, M. M. Fernández Sánchez & D. Gile (Eds.), *La evaluación de la calidad en interpretación: investigación* (pp. 109–123). Granada: Comares.

Hartley, A., Mason, I. Peng, G., & Perez, I. (2004). *Peer- and self-assessment in conference interpreter training.* Sponsored by CILT. Retrieved from http://www.lang.ltsn.ac.uk/prf.aspx#lang1.

Iglesias Fernández, E. (2006). La indefinición del parámetro "agradabilidad de la voz" y los estudios de calidad de la interpretación simultánea. In M. J. Varela (Ed.), *La evaluación en los estudios de traducción e interpretación* (pp. 225–239). Seville: Bienza.

Iglesias Fernández, E. (2007). La incidencia del parámetro agradabilidad de la voz. In A. Collados Aís, E. M. Pradas Macías, E. Stévaux & O. García Becerra

(Eds.), *La evaluación de la calidad en interpretación simultánea: parámetros de incidencia* (pp. 37–51). Granada: Comares.

Iglesias Fernández, E. (2010). Speaker fast tempo and its effects on interpreter performance: A pilot study of a multilingual interpreting corpus. *The International Journal of Translation, 22*, 205–228.

Kopczynski, A. (1994). Quality in conference interpreting: some pragmatic problems. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the gap. Empirical research in simultaneous interpretation* (pp. 87–99). Amsterdam/Philadelphia: John Benjamins.

Kurz, I. (1989). Conference interpreting: User expectations. In D. Hammond (Ed.), *Coming of age. Proceedings of the 30th Conference of the ATA* (pp. 143–148). Medford, N.J.: Learned Information.

Kurz, I. (1993). Conference interpretation: Expectations of different user groups. *The Interpreters' Newsletter, 3*, 13–21.

LNTO. (2001). *The National Standards in Interpreting*. London: Languages National Training Organisation.

Mack, G. & Cattaruzza, L. (1995). User surveys in SI: A means of learning about quality and/or raising some reasonable doubts." In J. Tommola (Ed.), *Topics in interpreting research* (pp. 37–49). Turku: University of Turku, Centre for Translation and Interpreting.

Mackintosh, J. (1995). A review of conference interpretation: Practice and training. *Target, 7*(1): 119–133.

Martin, A. & Abril Martí, M. I. (2003). Teaching interpreting: Some considerations on assessment. In A. Collados, M. M. Fernández, E. M. Pradas, C. Sánchez-Adams, E. Stévaux (Eds.), *La evaluación de la calidad en interpretación: docencia y profesión* (pp. 197–208). Granada: Comares.

Moser, P. (1995). *Expectations of users of conference interpretation*. Final report commissioned by AIIC. Vienna, SRZ Stadt und Regionalforschung GmbH.

Namy, C. (1978). Reflections on the training of simultaneous interpreters: A meta-linguistic approach. In D. Gerver & H. Sinaiko (Eds.), *Language, interpretation and communication* (pp. 343–352). New York: Plenum Press.

Pérez-Luzardo, J., Iglesias Fernández, E., Ivars, A. & Blanco Mayor, M. J. (2005). Presentación y discusión de algunos parámetros de investigación en la evaluación de la calidad en interpretación simultánea. In *Actas del II Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación* (pp. 1133–1154). Madrid. Universidad Pontificia de Comillas.

Pradas Macías, E. M. (2003). *Repercusión del intraparámetro pausas silenciosas en la fluidez: Influencia en las expectativas y en la evaluación de la calidad en interpretación simultánea*. Doctoral Dissertation, Universidad de Granada, Granada.

Pradas Macías, E. M. (2004). *La fluidez y sus pausas: Enfoque desde la interpretación de conferencias.* Granada: Comares.
Pradas Macías, E. M. (2006). Probing quality criteria in simultaneous interpreting: The role of silent pauses. *Interpreting, 8*(1), 25–43.
Riccardi, R. (2002). Evaluation in interpretation: Macrocriteria and microcriteria. In Hung, E. (Ed.), *Teaching translation and interpreting 4*, (pp. 115–126). Amsterdam/Philadelphia: John Benjamins.
Russo, M. C. (2005). Simultaneous film interpreting and users' feedback. *Interpreting, 7*(1), 1–26.
Schäffner, C. & Adab, B. (Eds.) (2000). *Developing translation competence*. Amsterdam/Philadelphia: John Benjamins.
Schjoldager, A. (1996). Assessment of simultaneous interpreting. In C. Dollerup & V. Appel (Eds.), *Teaching translation and interpreting 3. New horizons: Papers from the third language international conference* (pp. 187–195) Elsinore, Denmark, 9–11 June, 1995. Amsterdam/Philadelphia: Benjamins.
SCIC (2007) *Customer Satisfaction Survey* . Retrieved from http://ec.europa.eu/dgs/scic/contentu/customer_satisfaction_survey_en.htm
SCIC (2010) *Customer Satisfaction Survey* . Retreived from http://ec.europa.eu/dgs/scic/newsinitpage/index_en.htm
Seleskovitch, D. (1968). *L'Interprète dans les conférences internationales*, Paris, Minard.
Shlesinger, M. (1997). Quality in simultaneous interpreting. In Y. Gambier, D. Gile, & C. Taylor (Eds.), *Conference interpreting: Current trends in tesearch* (pp. 123–131). Amsterdam/Philadelphia: John Benjamins.
Vuorikoski, A. R. (1993). Simultaneous interpretation – user experience and expectations. In C. Picken (Ed.), *Translation-the vital link. Proceedings of the XIII. FIT World Congress* (pp. 317–327). London: ITI.

Rethinking Bifurcated Testing Models in the Court Interpreter Certification Process

Melissa Wallace[1]
University of Wisconsin – La Crosse

In the United States, with 44 out of 50 states holding membership in the Consortium for Language Access in the Courts, the court interpreting certification exam administered by this entity holds absolute primacy and is the most important gatekeeper to the profession. This study explores whether or not success in one mode of interpreting on the Consortium's oral certification exam could potentially predict successful performance in the other two modes; likewise, the viability of utilizing an abbreviated testing model, positing the mode that appears to predict overall success as a screening exercise, is contemplated. In order to isolate a potential predictor mode, this chapter explores precedents for the use of abbreviated testing models with Consortium exams and recreates a small study on a vastly larger scale, contributing an evidence-based analysis of some 6,000 raw exam scores spanning over a decade. With substantial data supporting the relationship between success in the simultaneous mode and overall success on the Consortium certification exam, the implementation of a bifurcated model could have a potentially very real impact on the way the Consortium exam is administered.

Key words: court interpreting, certification exams, assessment, performance, abbreviated testing models

1. Introduction

Interpreting studies stakeholders, from practicing professionals to researchers, express broad consensus on the significance of and justification for further inroads into empirical studies focusing on assessment and testing in interpreting studies. Assessment is vitally important not only for the purpose of screening applicants for entry into educational programs, providing feedback for students as they progress in their training, or testing their knowledge and skills at the end of a course of study, but most germane to the present discussion, it is essential for qualifying exams such as the certification exams used in the field of court interpreting. Whereas traditional aptitude testing mechanisms seek to identify interpreting candidates with the potential to acquire interpreting skills by the end of a course of study, qualifying exams identify candidates who are already able to meet a minimum standard in interpreting performance. Bontempo and Napier, scholars of sign language interpreting, are fully versed in recent scholarship regarding spoken language interpreting, and they emphasize that while

1 mwallace@uwlax.edu.

> … there appears to be general agreement about some of the skills needed in a candidate that may be assessable by an ability test at program admission (such as knowledge of working languages), less agreement and substantially less research supports factors of aptitude that may be predictive of interpreter performance. Which personality / affective factors (such as anxiety, motivation, stress-resistance, emotional sensitivity, and confidence, among others) and cognitive abilities (for example, intelligence, memory capacity, processing speed, attention span etc.) are predictive of individual performance…?… How exactly can aptitude for learning the complex skills required in interpreting be assessed in an efficient and effective manner…? (Bontempo & Napier, 2009, p. 251).

Because the majority of the current body of scholarship on interpreter ability treats recent studies on selecting apt students for training programs, such studies will inform the present discussion. Indeed, aptitude is closely linked to predictors of successful performance. In the United States, court interpreter certification is entirely performance-based[2], in contrast to some countries in which interpreters are considered qualified to practice in court based solely on success on written exams or completion of specific undergraduate degrees. The Federal Court Interpreter Certification Exam (FCICE), the National Judiciary Interpreter and Translator Certification (or NJITCE, which is the National Association of Judiciary Interpreter and Translator Association's certification for Spanish-language court interpreters and translators), as well as the certification exam used at the state level which is administered by the Consortium for Language Access in the Courts, all require their candidates to pass exercises in simultaneous interpreting, consecutive interpreting, and sight translation. The Consortium exam is the most commonly required credential in the country as it allows court interpreters to practice in municipal, county and state venues. Currently 44 out of 50 states hold membership in the Consortium for Language Access in the Courts; therefore, the court interpreting certification exam administered by this entity holds absolute primacy and is the most important gatekeeper to the profession.

This chapter explores whether or not success in one mode of interpreting on the Consortium's oral certification exam could potentially predict successful performance in the other two modes. Likewise, the viability of utilizing an abbreviated testing model, using the mode that appears to predict overall success as a screening exercise, is contemplated. In order to isolate a potential predictor mode, precedents

2 Hildegard Vermeiren, Jan Van Gucht and Leentje De Bontridder contrast social interpreter testing in Flanders, Belgium, where candidates are tested strictly on competencies relevant to interpreter performance, with testing in other countries such as Sweden or the United Kingdom, which assess interpreting performance as well as cultural and terminological knowledge (2009, p. 307).

for using abbreviated exam models by Consortium member states will be discussed, followed by an evidence-based analysis of some 6,000 raw oral exam scores spanning over a decade. By closely scrutinizing the empirical results reflected in small but significant field studies as well as the analysis of the Consortium data set, patterns of performance begin to emerge which may aid researchers in determining whether or not successful performance in one mode of interpreting can predict success in other modes. Unlike purely theoretical studies, the results contained herein are based almost exclusively on publicly available and internal reports published by industry stakeholders at the Consortium for Language Access in the Courts, in addition to personal interviews and communications with former and current court interpreting program managers in several key US states. The data-driven result is the consideration of the implications of a bifurcated testing model which posits the simultaneous mode as a predictor of successful performance on Consortium exams. With substantial data supporting the relationship between success in the simultaneous mode and overall success on Consortium certification exams, the implementation of a bifurcated model could have a potentially very real impact on the way the Consortium exam is implemented and how future test development resources are allocated, especially as related to languages of lesser diffusion.

2. Precedents for abbreviated testing models in court interpreter certification testing

The theoretical implications of isolating and identifying a predictor mode for court interpreting success are abundant; but what are the practical benefits? The search for a predictor mode on the Consortium exam, as it turns out, was born directly from the desire to solve a series of practical concerns faced by court interpreting program managers who believed that the implementation of an abbreviated test model could alleviate a series of logistical problems. To that end, before initiating an analysis of test score-related data, let us discuss the light that this data can shed on positing a bifurcated method as an alternative model for court interpreter certification testing. Beginning with a brief definition of bifurcated testing models, the potential desirability of an abbreviated testing model shall be explained in the context of the pragmatic impetuses behind one state's statistics-based search for a predictor mode. This section will also describe the subsequent involvement of the Consortium's Technical Committee in considering abbreviated testing models[3], discuss the realities of the three US states that

3 The Technical Committee of the Consortium for Language Access in the Courts is responsible for the construction and design of court interpreter performance exams, the administration of

currently utilize a bifurcated model, and consider the implications of the use of a bifurcated model in relation to testing languages of lesser diffusion as well as the staffing and cost-reduction strategies associated with its use.

2.1 A practical impetus for exploring abbreviated test models: The case of New Jersey

By way of definition, a bifurcated certification testing method tests simultaneous interpreting, consecutive interpreting and sight translation exactly as a traditional oral certification does, with the difference that it simply does it in two phases. In other words, one of the exercises testing a specific mode of interpreting is used as a screening exercise, and candidates who pass it are then allowed to sit for the other two exercises. In the three US states which currently use a bifurcated testing method, all three utilize the simultaneous mode as an initial screening instrument. Candidates who pass the simultaneous exercise are then allowed to take exams in consecutive interpreting and sight translation.

Adoption of the bifurcated approach to testing found its genesis in the state of New Jersey. According to Robert Joe Lee, former New Jersey program manager and former voting member of the Technical Committee[4], as resources began to dwindle in the 1990s, finding a more cost-effective way of identifying competent interpreters became an economic necessity as well as a pragmatic decision in his state. Lee had serious concerns about using taxpayers' money responsibly at a time when New Jersey did not charge its prospective interpreters for testing, bearing the cost completely at the state level. Costs related to testing and rating became unsustainable. In order to pare back expenses, New Jersey began by eliminating one of its three qualified raters who, at that time, were also the test administrators[5]. The state had been paying two raters $250 a day to administer and rate six exams per day and, according to Lee, so many of the exams

court interpreter exams (including instructions), the content of test construction manuals, the rating of court interpreter exams, language-specific exam development, establishing recommended qualification standard levels for interpreters, establishing minimum qualifications for test administration and rating, assessment of tests developed by the Consortium, and assessment of tests administered by other organizations (Technical Committee, 2009, p. 1).

4 Many thanks are due to Robert Joe Lee for sharing his expertise, data, and time over the course of many invigorating email and telephone conversations.

5 The three raters were comprised of one academic linguist and two practicing interpreters with federal certification. When the decision was made to eliminate one of the raters, instead of suppressing one of the "categories" (i.e. academic versus certified practitioner), Lee looked for diversity of language background, training and expertise (R. J. Lee, personal communication, March 14, 2011).

administered revealed such vast incompetency that it did not seem sustainable to continue to offer the entire exam, at such a high expense, when so few qualified interpreters were being identified. Lee in New Jersey as well as program managers from other states sought to remedy this cost-benefit dilemma and search for a way to screen test-takers more effectively (R. J. Lee, personal communication, March 14, 2011).

Under Lee's direction, the New Jersey office endeavored to carry out a series of internal studies on which to base decisions regarding the use of abbreviated testing models. The New Jersey Administrative Office of the Courts began by conducting systematic time studies, isolating the mean time per assignment for which each mode of interpreting accounted. In the first time study which took place during the weeks of June 7–11 and 14–18 of 1993, all interpreted events served by full-time interpreters statewide in the Superior Court were analyzed (Technical Committee, 2001, p. 15). Simultaneous interpreting accounted for 66 % of the time used during the assignments measured, followed by consecutive at 57 %. Simultaneous occurred more than any other function measured in the average assignment and also lasted a longer mean time. Sight translation, at 22 %, seemed comparatively negligible in frequency, taking the least average amount of time per assignment of all interpreting modes. Time study two, conducted using the same variables during the weeks of March 11–15 and 18–22 of 1996, shows a reversal of time spent per function as the consecutive mode (in which 76 % of assignments took place) then appeared in a greater proportion of assignments and lasted longer than simultaneous at 69 % (Technical Committee, 2001, p. 16). This second set of results, in combination with the consecutive mode's strong second-place showing in the first study, begs the question: why isolate simultaneous as significantly more essential in forming part of an abbreviated test model? Interpreting is a very complex task, say Timarová and Ungoed-Thomas, and "it is not reasonable to expect that one supertask will be found which will serve as the sole predictor" of interpreter ability (2009, p. 242). Is the simultaneous mode the "supertask" which can be looked to as a beacon of predictive validity in identifying qualified interpreters? Would the mix of tasks inherent in this one mode of interpreting and their discrete isolation and identification be useful to consider when designing and interpreting summative exams?

Both sets of results show that simultaneous is either the most frequently used mode in the assignments contemplated in the studies or that it is in a close second with consecutive, but what does frequency really reveal in terms of a mode's importance or ability to predict performance in other modes? In order to answer this question, New Jersey undertook an additional study based on the systematic isolation of the three interpreting modes in order to examine the impact of passing each mode on passing the remaining parts of the test. Until now, these findings

constituted the only systematic attempts made at specifically examining modes of interpreting as predictors of performance on the Consortium certification exam, and are thus especially relevant to the questions under consideration herein.

The languages measured were Haitian Creole, Polish, Portuguese and Spanish, with a total of 134 candidates for the consecutive and simultaneous exercises, and 126 candidates for the sight translation component, as New Jersey at the time did not offer a sight translation component for Haitian Creole. As evidenced in Table 4, when isolating for success on the sight translation component and its impact on passing the other two modes, 33 % of all examinees who passed the sight translation exercise also passed the other two modes. The consecutive mode's correlation to the passing of the other two modes was 51 %. As regards the simultaneous mode, 81 % of candidates who passed the simultaneous exercise also passed the other two components (Technical Committee, 2001, pp. 16–17). But what level of prediction would program managers and Consortium officials consider reliable enough in terms of test validity in order to contemplate the use of abbreviated test models on a larger scale? According to the above results, there exists a full 48 % difference between the success of simultaneous passers and sight translation passers in achieving success on both of the other two modes. Would these numbers vary significantly if a larger sample were used? Are they high enough to consider consecutive more seriously as a contender in an abbreviated test model? What can be safely affirmed is that sight translation seems to be the mode least likely to predict passing-level performance in comparison to simultaneous and consecutive.

2.2 Exploring abbreviated testing models at the Consortium level: The Technical Committee

An absorbing report published in 2001 by the Technical Committee of what was then called the Consortium for State Court Interpreter Certification relies quite substantially on New Jersey's quantitative studies, the first systematic and data-driven explorations of the viability of abbreviated testing models in the context of Consortium court interpreting certification. The genesis of the exploration of abbreviated models is thus: the Technical Committee, comprised of voting members from the states of New Jersey, Minnesota, Florida, Massachusetts and California, along with three other non-voting members, acknowledged a series of challenges in identifying and certifying qualified interpreters in a reliable and cost-effective way. The Committee set about to address them systematically through a contemplation of abbreviated testing models and a statistical justification for modifying the traditional oral exam comprised of exercises in all three

modes of interpreting for certain languages for which standard test models did not exist. The report begins by describing the dilemma faced by the Consortium at the time. They acknowledged that

> while in theory it is desirable to write a standard performance test (i.e., two sight components [one in each direction], consecutive, and simultaneous … to certify court interpreters in all languages desired by member states, the reality is that this is not presently and probably never will be feasible (Technical Committee, 2001, p. 1).

Moreover, conversations among the states regarding the selection of additional languages for which to write new tests had begun to break down for the first time; the cost of adding tests for new languages to the Consortium's test bank hovered between $25,000 and $35,000 (Technical Committee, 2001, p. 1). Economic concerns, however, were not the only ones. Issues of access as well as the way interpreters of untested languages were perceived and compensated were also important considerations as the Technical Committee recognized that interpreters working in untested languages

> do not benefit from the professionalization that interpreters who work in languages that are tested enjoy. This includes lack of both professional recognition (they are often labeled by terms such as 'otherwise qualified') and professional treatment (they are often paid substantially less than 'fully' certified interpreters) (Technical Committee, 2001, p. 1).

Holding as one objective, then, that of being more inclusive in terms of language representation options in the test bank, the Technical Committee put forward the question of an alternative test model as an acceptable solution, and undertook the exploration of this question as a priority. Although the committee's report clearly acknowledged the widely-held belief that a certification test must be valid and reliable if it is to be depended upon to identify qualified court interpreters, they also recognized that

> … any test that can weed out people who cannot demonstrate ability to perform at least some of the highly sophisticated modes of interpreting would be welcome… It is better to know for sure that an interpreter can perform even one mode of interpreting, than not to know it (Technical Committee, 2001, p. 2).

While the report explores and contemplates three different proposed abbreviated models, the Technical Committee agreed upon three basic principles to guide its review of possible models and ultimate recommendation to the Consortium. They agreed that any model ultimately adopted by the Consortium should have three characteristics: it should include at least the simultaneous mode, it should predict ability to perform as many modes of interpreting as possible which are not directly included in the model, and it should be easy and cost-effective to both develop and administer. The Committee felt that the inclusion of the simultaneous mode was germane to an abbreviated test based on the previously discussed

New Jersey statistics, affirming that passing the simultaneous exam could be taken as a "surrogate indicator that (examinees) have a high probability of passing sight and consecutive if those components were available in the language" (Technical Committee, 2001, p. 10). The abbreviated model ultimately recommended consisted of the use of the simultaneous exercise as a qualifying exam, along with a "conversational piece[6]" aimed at having some basis on which to assess English proficiency as the simultaneous exercise only includes English-to-foreign language production. The advantages of the proposed abbreviated model were that it was the option that was by far the least expensive model to develop and administer, as probably any existing Consortium simultaneous test could be taken and revised to include the appropriate distribution of scoring units. At that point, the main expense to the Consortium would be the recruiting and training of examiners who could develop the dictionaries of acceptable and unacceptable renderings and do the grading. The report adds, "This test could be even administered before there is an assembled team of raters since it could be proctored to any interpreter of ANY language" (Technical Committee, 2001, pp. 6–7).

2.3 The bifurcated method in practice: New Jersey, New Mexico and Idaho

Despite the obvious appeal of a shorter test which is easier to administer, it is important to acknowledge the potential for and justification of resistance to bifurcated or other abbreviated exam models, if for no reason other than the importance of task validity as a psychometric testing principle. In testing theory, an assessment tool only has task validity if it asks examinees to perform the tasks that the exam is meant to measure. There does not appear to be consensus among Consortium states about use of the bifurcated approach and, indeed, most states require that all three exercises of the oral exam be administered all in one sitting in order to afford the candidate the opportunity to demonstrate that he or she has the requisite stamina to perform all three skills in the allotted time.

With this understanding, an attempt was made to speak with current program managers of the three states which currently use the bifurcated testing method. In New Jersey, Manager of the Language Services Section of the Administrative Office of the Courts, Brenda Carrasquillo, chooses to continue to administer certification exam in two phases based on the extensive analysis of the bifurcated approach carried out by her predecessor, Robert Joe Lee. Ms. Carrasquillo

6 The conversational piece was meant to identify consistent non key-word issues such as problems with pronunciation and fluency, although there were concerns that it could be used to override an objective score, at the same time as it introduced the possibility of rater bias and prejudice. The conversational piece was discontinued eventually.

alludes to the practical concerns addressed by the bifurcated method in stating that "Part of the reason we choose to administer the exam in two phases is simply due to AOC staffing issues since our office is very small" (B. Carrasquillo, personal communication, February 24, 2011). In New Mexico, Statewide Program Manager at the Administrative Office of the Courts Pamela Sánchez explained that candidates "take the consecutive and sight within a year of passing the simultaneous examination" (P. Sánchez, personal communication, February 15, 2011), a fact which accounts for the testing results published in New Mexico's year-end report to the Consortium. For 2010, a year in which all test-takers took Spanish-language exams, testing totals were as follows[7]:

Table 1. New Mexico 2010 testing results

Exam	**Number of candidates tested**	**Number of successful candidates**	**Pass rate**
Simultaneous	45	15	33 %
Consecutive	29	10	34 %
Sight translation	29	10	34 %
Overall passing rate	45	10	22 %

(adapted from Sánchez, 2010, p. 4)

Without demographic information about the examinees[8], it is difficult to compare the 33 % pass rate in the simultaneous mode to other data sets, although the pattern that clearly emerges from New Mexico's annual report is that the number of test-takers handled by the office is significantly reduced by using simultaneous as a screener.

Janica Bisharat, Program Manager of the Administrative Office of the Courts in Idaho, provided some very illuminating reflections on their office's use of the bifurcated approach. She states,

> Our reason for using the bifurcated testing approach is simply because we have very limited resources. We have tried both methods of testing and have found the bifurcated approach to be more efficient and quite frankly easier on our test proctor. We have one staff member who checks in exam candidates and proctors all the exams. Administering the oral exam in one

7 Presumably 14 of the 29 test-takers in consecutive and sight translation had passed the simultaneous exam in 2009.

8 "… the Consortium doesn't conduct any type of data collection of the test candidates; very briefly there were attempts to collect information from candidates about whether or not using the Practice Exam Kit was helpful preparation for the exam but even that was not very successful and we have since stopped asking for it" (C. Green, personal communication, February 11, 2011).

> sitting to a number of candidates over the course of a couple of days can be pretty taxing. We want to ensure that each exam is administered according to policy and without incident or disruption. It is our experience that we don't administer nearly as many exams using the bifurcated approach, but we have not formally gathered any data to support this view (J. Bisharat, personal communication, April 21, 2011).

The Technical Committee's findings and recommendations, as previously discussed, strongly hint at the probability of the simultaneous mode as a predictor of aptitude. Presumably New Jersey, New Mexico and Idaho, in their efforts to optimize limited resources, feel justified in using the simultaneous mode as the screening part of the bifurcated approach. More research is clearly warranted, however. Recall that no studies similar to those carried out in New Jersey had ever been reproduced in any context, in any state, in any language pair, until similar parameters were applied to a vast data set of some 6,000 Spanish / English test scores provided by the Consortium for Language Access in the Courts.

3. Consortium data and the potential predictive validity of performance in the simultaneous mode

Indeed, while the smaller-scale predictor mode studies conducted in New Jersey seemed to handily pose simultaneous as a predictor of success on the rest of the oral exam, a rare opportunity arose to reproduce the study on a vastly larger scale. In search of more data which could prove or disprove the Technical Committee's hypothesis, contact was made with the Consortium for Language Access in the Courts in an effort in inquire about obtaining raw test scores in order to discern what the numbers would communicate in a broader quantitative analysis. Carola Green, Coordinator for Court Interpreting Testing Services and Operations at the National Center for State Courts, provided data consisting of exam scores for the Spanish / English language pair dating back from 1995 through 2009[9]. These several hundred pages of raw test scores, devoid of all identifying information, reflect the performance of test takers who took all three portions of the oral exam in one sitting, and promised to illuminate performance patterns of Spanish / English test takers.

While the relationships between the numbers can be examined from myriad perspectives, this chapter's analysis of the data is modeled in part on the New Jersey predictor mode study which posits successful performance in one mode of interpreting as a potential predictor of success in the other two modes for Con-

9 The author is deeply indebted to Carola Green and to the Consortium for Language Access in the Courts for sharing such valuable data. This gesture of good will and collaboration between official organisms and the academic world sets a laudable example for stakeholders engaged in the research, training and professional practice of interpreting.

sortium oral exam purposes. The following table provides a descriptive overview of the data analyzed[10]:

Table 2. Descriptive statistics of the Consortium data set[11]

	N	Minimum	Maximum	Mean	Standard deviation
ST score Eng > Span	5916	-1	29	15.65	4.514
ST score Span > Eng	5916	-1	29	15.21	4.658
Sight combined	5916	-2	53	30.87	8.380
Consecutive score	5916	-1	88	50.14	11.748
Simultaneous score	5916	-1	77[11]	41.83	13.573

Although no personal or demographic information is collected by the Consortium or associated with exam scores, an important amount of data is available for analysis. The Consortium data set comprises the total number of candidates (5,916) in the Spanish / English language pair who, from 1995 to 2009, took the certification exam in states which administer all three parts in one sitting. Even in the absence of personal information corresponding to the examinees, a considerable number of interesting patterns can be gleaned simply by knowing the maximum number of points based on correct responses in addition to the passing cut-off score.

While each member state can exercise its right to establish cut-off scores and requirements for certification, (for example, in some states candidates have to achieve a *combined* score of 70 % in order to pass the two sight translation exercises), the overwhelming majority of states conform to standard Consortium pass standards; that is to say, they require a 70 % score or better on all three scorable exercises in order to earn certification. A regards the standard required to pass a Consortium oral exam,

> the cut-score was determined by the Consortium Technical Committee based on the Federal court interpreter oral exam cut-score and research conducted by the state of New Jersey during its original program implementation. Specific data considered when evaluating the performance of an oral exam consists solely of an objective assessment and the cut-score percentage assigned when based on the scoring units rendered correctly by a candidate (ALTA Language Services, Inc., 2010, p. 15).

10 The author wishes to thank the Statistical Consulting Center at the University of Wisconsin – La Crosse for its assistance with statistical calculations of the Consortium data set, most especially Dr. Sherwin Toribio. Any errors of fact or interpretation remain the sole responsibility of the author.

11 An apparent anomaly, one examinee was awarded a score of 77 on the simultaneous exercise, for which 75 points is the maximum score.

Other tangentially related studies, furthermore, suggest that the choice of 70% is sound, such as Stansfield and Hewitt's examination of the predictive validity of written screening tests for court interpreters, which argues that a 75% cut-off on the federal oral exam is as equally justifiable on a statistical level as the current cut-off of 80% (Stansfield & Hewitt, 2005). In another study on interpreter aptitude tests used to identify candidates likely to be successful in a program of study, the author found the 70% cut-off rate to be reasonable and appropriate, stating that "the requirement of 70% acceptable answers set as the limit for passing the written pre-test is a relatively good indicator of the students' interpreter potential. That is to say, those who did *not* meet this limit also were *not* able to improve their interpreting performance during this short (14-week) training course to an extent that they reached an acceptable level of interpreting quality" (Skaaden, 1995, p. 95). While the aforementioned contexts differ from a credentialing exam, a 70% cut-off was deemed an appropriate pass rate for the analysis of the Consortium data precisely because it is the standard embraced almost universally by member states.

Furthermore, it is worthy of mention that a minimum score of -1 is possible on each exercise of the oral certification exam if the candidate does not attempt to take it at all, thus explaining the possible score of -2 on the sight translation mode, which is comprised of exercises from English to Spanish and vice versa. It should also be noted that several exam versions for Spanish are used, all of which are normed according to guidelines in the Oral Exam Construction manuals. Additionally, at some point the exam format changed to include fifteen more points on the consecutive exercises[12]. In analyzing the data set, a variable was added to allow for filtering of the two exam formats. The results discussed in this chapter, however, make conclusions across the two formats, as the same parameters of analysis were applied to each one. Furthermore, the fact that the Consortium data set represents an entire population and not a mere sample means that the data is not subject to sample variability and that the numbers discussed are actual values.

How, then, do overall results from the Consortium data set compare to the New Jersey study in terms of isolating a potential predictor mode of success on the overall oral certification exam?

12 The author was not able to learn when the Consortium augmented the consecutive exercises in Spanish nor what the rationale was. In changing the test from a total of 200 to 215 points, the overall pass rate across the 5916 scores in the data set provided went from 20% for the 200 point version to 16% for the 215 point version, representing a 20% drop in pass rates.

Table 3. Consortium data set: Modes as predictors of success

Predictor mode	# of examinees passing mode	% of examinees passing entire exam
Sight translation	2,331	45.6 %
Consecutive	1,758	60 %
Simultaneous	1,530	69 %

Table 3 depicts these results. In short, out of the 5,916 examinees who took the Spanish / English test, 2,331 passed the sight translation part by scoring a 70 % or better in each language direction. Out of these 2,331, only 1,062, or 45.6 %, passed the entire test by scoring a 70 % or better in each of the three modes of interpreting. 1,758 examinees out of 5,916 passed the consecutive exercise; out of these, 1,062 passed the entire test. 60 % of consecutive passers were able to pass the entire exam. Out of 5,916, 1,530 people passed the simultaneous exercise, with 69 % of them passing the entire oral certification exam. The following table presents a comparison of the results of the two studies:

Table 4. New Jersey study and Consortium data set: Modes as predictors of success

Predictor mode[13]	% of examinees passing entire exam (New Jersey study)	% of examinees passing entire exam (Consortium)
Sight translation	33 %	45.6 %
Consecutive	51 %	60 %
Simultaneous	81 %	69 %

The two data sets examined in this chapter differ in terms of the languages tested, their geographic scope (one state versus all member states), and the sheer number of examinees contemplated in the analyses. Similarities emerge, however, which suggest that of the three modes of interpreting, simultaneous is significantly more effective as a predictor of success on the other two modes, followed by consecutive and sight translation in the same order on both analyses. While the percentages differ, interpreting stakeholders and Consortium member state program managers may benefit from the results generated in order to empirically inform their decisions about using or rejecting abbreviated testing models.

13 To clarify, examinees who passed the predictor mode in question also passed the other two modes on the oral certification exam.

4. Conclusions, recommendations and avenues for further inquiry

The wealth of information provided in the Consortium data set cannot be underestimated, and the possibilities for its analysis are virtually limitless. Nevertheless, the scope of this study is limited to examining individual modes of interpreting as potential predictors of success on the entire oral certification exam, as well as to contemplating the potential for utilizing such information in the context of interpreter certification testing. The case studies and data considered in this chapter, then, pave the way for several conclusions, generate some as yet unanswered questions, and offer avenues for further inquiry. Examination of the data suggests that there are several key questions or areas of consideration which Consortium and other official certification testing organizations may want to consider.

First, the predictive validity of using simultaneous as a predictor mode, at 69 % – 81 %, is convincingly high, especially when we consider that 70 % is the minimum standard cut-off rate recommended by the Consortium. Upon considering the Consortium data set which measures only Spanish / English scores, even if this population of test-takers has specific features which cannot be extrapolated to apply to speakers of other Consortium languages, member states may benefit from weighing the cost-saving benefits of using the bifurcated method in administering the Spanish test, using simultaneous first as a preliminary screening instrument. Of vital importance to future research in this vein would be the possibility of cross-referencing demographic data with raw test scores. If the Consortium and other similar testing bodies began to collect and reference important identifiers such as gender, native language, level of education, etc., it would prove illuminating in our attempt to draw reliable profiles of interpreters with the requisite skills to pass the oral certification exam.

Naturally, the implications of this study extend beyond the confines of institutional and professional arenas into the areas of education and training. With hard data confirming that the interpreters who possess the skills to pass simultaneous exercises have statistically significantly higher chances to pass consecutive and sight translation exercises, university or training programs would be justified in using entry or aptitude exams which test for the inherent skills involved in simultaneous interpreting.

Other questions to consider include that of lending due consideration to the consecutive mode, given that 60 % of Spanish / English candidates who passed consecutive also passed the other two exercises. Is this level of predictability convincing or valuable? Traditionally the consecutive mode has been considered to be the most vital in the court interpreting process, although the reality in municipal, state, and county courtrooms may be otherwise. More data is needed in order

to re-evaluate the importance of training and testing in the consecutive mode in the court interpreter context in addition to that of simultaneous interpreting.

Furthermore, in contemplating the use of a bifurcated testing model, official credentialing organizations must not lose sight of the necessity of adequately upholding the highest standards of test validity (including task validity) in order to ensure that court interpreter certification continues to be valid and reliable. Interpreting stakeholders must acknowledge traditional institutional and professional resistance to abbreviated testing models. Currently 41 out of the 44 Consortium member states require their candidates for whom full exam versions exist to test in all three modes of interpreting at the same time. Testing bodies must decide whether or not predictor mode studies are convincing enough in order to contemplate using a simultaneous exercise as a preliminary screening exam, with the consecutive and sight translation portions to be administered later if they exist. These concerns should also be balanced by a realistic and data-driven analysis of the real cost-saving measures enjoyed by states using the bifurcated approach; in other words, the perceived benefits of savings and use of personnel should be scrutinized and weighed in order to determine whether or not the benefits constitute a possible model that could or should be imported to other states.

Finally, and of utmost concern among Consortium member states as well as European Union nations, the results of this study should inform stakeholders who care about testing and certifying interpreters in languages of lesser diffusion. Even though a full exam may be most desirable, the reality is that even abbreviated exams do not exist for a plethora of languages. With exam development costs soaring and ubiquitous budget constraints, some states have turned to using Oral Proficiency Interviews, or OPIs, to assess a candidate's proficiency in the non-English language by measuring his or her ability to use the language effectively and appropriately in real-life situations. The obvious failing of a foreign language-only OPI is that it does not at all measure skills in the language of record nor, most essentially, does it measure interpreting skills. Perhaps the Consortium would do well to consider allocating funds for developing abbreviated versions of simultaneous exercises in the most critical languages for which full exam versions have not yet been developed. In essence, the Consortium Technical Committee should consider rejecting the use of only written tests or OPIs as minimally qualifying exams in favor of investing in the elaboration of simultaneous exercises in languages of lesser diffusion. While acknowledging that test development resources are hanging in a critical balance, perhaps the Consortium would do well to consider allocating funds for developing abbreviated exam versions (consisting of simultaneous exercises) in the most critical languages of lesser diffusion for which full exam versions have not yet been developed.

In sum, in a profession which struggles constantly with restraints on economic and human capital, further examination of specific modes of interpreting as predictors of global certification exam success constitute an important albeit preliminary contribution to the field of certification exam research.

References

ALTA Language Services, Inc. (2010). *Beyond words: California assessment of the Consortium for Language Access in the Courts' exams*. Retrieved August 14, 2011, from Judicial Council of California, Administrative Office of the Courts website: http://www.courts.ca.gov/documents/ALTAReport.pdf

Bontempo, K., & Napier, J. (2009). Getting it right from the start: Program admission testing of signed language interpreters. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice* (pp. 247–295). Amsterdam: John Benjamins.

Consortium for state court interpreter certification survey: Certification requirements 2010 (2010, March 31). Retrieved from http://www.ncsconline.org

Hewitt, W. (n.d.). *Expert panel on community interpreter testing and certification: Final report*. Retrieved from http://www.matiata.org

Hewitt, W. E. (n.d.). Court interpreting proficiency tests: A summary of what they look like and how they are developed. *National Center for State Courts*. Retrieved from http://www.ncsconline.org

Improving interpretation in Wisconsin's courts: A report on court-related interpreting and translation with recommendations on statute and rule changes, budget items, interpreter training programs and certification tests, and judicial and professional education programs. (October 2000). Committee to Improve Interpreting & Translation in the Wisconsin Courts Report to the Director of State Courts.

Lee, R. (2000, April 7). *First thoughts regarding a written test for court interpreter certification*. Retrieved from http://www.ncsconline.org

Overview of the oral performance examination for prospective court interpreters (pp. 1–43, Publication). (2000). Consortium for State Court Interpreter Certification.

Rejsková, J. (1999). Establishing a correlation between performance in consecutive interpreting and potentially good performance in simultaneous interpreting. *Folia Translatologica*, *6*, 41–60.

Sánchez, P. (n.d.). *2010 Annual report* (pp. 1–5, Rep.). New Mexico Court Interpreting Program.

Skaaden, H. (1999). Lexical knowledge and interpreter aptitude. *International Journal of Applied Linguistics*, *9*(1), 77–97.

Stansfield, C. W., & Hewitt, W. E. (2005). Examining the predictive validity of a screening test for court interpreters. *Language Testing*, *22*(4), 438–462.

Technical Committee of the Consortium for State Court Interpreter Certification. *The Role of the Technical Committee*. Retrieved from http://www.ncsconline.org

Technical Committee of the Consortium for Language Access in the Courts. (2010, October). *Court Interpreter Oral Examination: Test Rating Manual* . Retrieved from http://www.ncsconline.org

Technical Committee of the Consortium for State Court Interpreter Certification. (2001). *Abbreviated test models*. Retrieved from http://www.ncsconline.org

Test rating standards and resource materials for rater training: Court interpreting oral proficiency examination (2001, June). Retrieved from http://www.ncsconline.org

Timarová, S., & Ungoed-Thomas, H. (2009). The predictive validity of admissions tests for conference interpreting courses in Europe: A case study. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting: A call for dialogue between research and practice* (Vol. XIV, American Translators Association Scholarly Monograph Series, pp. 225–245). Amsterdam: John Benjamins.

Vermeiren, H., Van Gucht, J., & De Bontridder, L. (2009). Standards as critical success factors in assessment: Certifying social interpreters in Flanders, Belgium. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting: A call for dialogue between research and practice* (Vol. XIV, American Translators Association Scholarly Monograph Series, pp. 297–328). Amsterdam: John Benjamins.

Assessing the Impact of Text Length on Consecutive Interpreting

Jungyoon Choi[1]
La Rochelle University, France

Multiple variables can affect the performance outcome in interpreting. Such variables can, either directly or indirectly, determine the difficulty level of the original text or source text. Against this backdrop, the length of the original text, which could be assumed to be one of the decisive factors in judging the difficulty level of a source text, could influence the performance outcome. To bring this matter to the forefront, this chapter will introduce source text variables, elaborate how source text variables such as text length could determine the difficulty of a source text, and further shed light on its possible influence on interpreting performance. Based upon the hypothesis that longer texts are likely to require more concentration and endurance for the interpreter than shorter ones, this chapter attempts to demonstrate how text length can potentially impact performance in consecutive interpreting through an experiment conducted on future interpreters at the graduate level. The performance will be assessed by a performance assessment tool developed by the author. The assessment results will be analyzed through a statistical tool to explore the possibility of text length influencing performance and empirically verify how text length can potentially affect the performance of would-be interpreters.

Key words: performance assessment, interpreting, source text variables, text length, endurance.

1. Introduction

1.1 Definition of source text variables

One of the major challenges that professional interpreters have to face is whether he or she is able to cope with the different variables that occur in various conference settings. As a result, an interpreter's performance can differ from one situation to another for many reasons. Due to the nature of interpreting, Choi (2008) asserts that performance can be influenced not only by variables outside the source text, often referred to as the original text, but also by elements that are directly related to the source text itself (p. 110). The former can be explained by the physical situation in which the speaker and interpreter are situated, and the interpreter's own ability (Choi, 2008, p. 111). For instance, this relates to the availability of texts (Kalina, 2002), work load and the interpreter's experience (Moser-Mercer, 1996), sound quality and visibility of the speaker (Altman, 1990), language combination and language direction (Moser-Mercer, 1996), stress endurance level, etc.

1 jungychoi@gmail.com

On the other hand, elements that are directly related to the source text itself correspond to what Choi (2008) refers to as *text difficulty*. This can also be explained by delivery speed and speech quality (Altman, 1990), text length (Nord, 1991), speech style (Seleskovitch & Lederer, 1995), whether the speaker is speaking in his or her mother tongue (Kalina, 2002) and so forth. Such variables can be defined as source text variables considering that they are in some cases unpredictable and can often influence the performance outcome. Such source text variables and variables outside the source text make the performance assessment in interpreting all the more complex particularly for researchers in this field.

1.2 Research objective

Not much has been said in detail as to how certain source text variables can influence interpreting performance. This could be, in part, attributable to the fact that researchers in the interpreting field, unlike translation, are required to overcome certain obstacles. In this regard, Choi (2008) has already pointed out the three major challenges that a researcher has to address in interpreting; data availability, comparability and unpredictability. First, data availability concerns recording constraints and confidentiality issues. Second, comparability relates to the difficulty of collecting the interpreted version from identical source texts. Third, unpredictability has to do with the multiple variables that an interpreter has to face during meetings (Choi, 2008, p. 42). Despite recent studies (Angelelli & Jacobson, 2009; Shelesinger & Pöchhacker, 2011) which have pointed to a wide range of different assessment issues in interpreting and translation, the aforementioned challenges have still made the assessment process all the more complex especially for interpreting in terms of reliability and validity.

Nevertheless, the objective of this research is to observe the possible trends of interpretation performance when source text variables such as text length vary in consecutive interpreting and, thus infer whether text length can affect the performance of novice learners. To that end, it should be noted that this research does not consist in assessing the performance per se but in observing how certain source text variables such as text length can possibly affect student performance.

2. Research method

2.1 Hypothesis

The experiment was based upon the hypothesis that, compared to shorter source texts, longer texts will probably require more endurance for novice learners, and thus affect their performance in consecutive interpreting. The source text uttered in the midterm test was approximately 50 seconds longer than the other two texts used respectively during the initial and final tests provided that the utterance speed of all three source texts respectively used in the initial, midterm and final tests approximates what Seleskovitch and Lederer (1995) refer to as normal speed range: 120 to 220 words per minute.

2.2 Test subjects

The text subjects were a total of 11 first-year master students in their second semester enrolled at the Graduate School of Interpretation and Translation (GSIT) at Hankuk University of Foreign Studies in Seoul, South Korea. Their average age was twenty-five years old and their language combination was Korean (A language) and English (B language). The course concerned was English into Korean consecutive interpreting. The students had already received 32 hours of training in English into Korean consecutive interpreting during their first semester prior to the experiment.

2.3 Test materials

Three tests were carried out to verify the hypothesis. An initial test took place during the third week of training of the 2nd semester. A midterm test took place approximately 6 weeks after the initial test. The final test also took place 6 weeks after the midterm test. For each test, a different source text in English was used. The texts were not too complicated or technical considering that the test subjects were novice learners who were still in their first year of learning consecutive interpreting.

The source texts uttered in the initial, midterm and final tests respectively lasted 3 minutes and 3 seconds for the initial test, 3 minutes and 52 seconds for the midterm test, and 3 minutes and 2 seconds for the final test. That is, as previously mentioned in the hypothesis, the source text uttered in the midterm test

was approximately 50 seconds longer than the other two texts used respectively during the initial and final tests.

The utterance speed for each text stood at 118 words per minute for the initial text, 129 words per minute for the midterm test, and 116 words per minute for the final test. The utterance speed of the three source texts ranging from 116 words per minute to 129 words per minute, approximated what Seleskovitch and Lederer (1995) referred to as normal speed range: 120 to 220 words per minute. All student performances for all three tests were recorded and assessed by two professional interpreters.

2.4 Variables

The independent variable used in the experiment was the length of the source text. As previously mentioned, among the three different source texts, the text used in the midterm test was approximately 50 seconds longer than the texts used in the initial and final tests. The dependent variables were the scores of overall performance, accuracy, expression and presentation. Explanation on these variables will be provided in the following sections in more detail.

2.5 Performance assessment

In order to assess the students' performance, the PAT (Choi, 2005), a performance assessment tool developed by the author, was adapted to be used to quantify their scores in terms of overall performance, accuracy, expression and presentation. To facilitate the assessment process, only the first 2 minutes and 30 seconds of the source text was assessed by the two raters, who were two professional interpreters who had the same language combination as the students: Korean as their mother tongue, and English as their B language or active language. They assessed the same group of students by utilizing the PAT. The average scores between the two raters were used for the statistical analysis. Following is a sample of the performance assessment results[2].

Adapted from the Performance Assessment Tool (PAT) developed by the author, the assessment criteria is classified into three major categories: *accuracy*, *expression* and *presentation*. Under each category, there are subcategories so as to

2 The abbreviations indicated in the PAT above are as follows (Choi, 2005); ER: Error range, OPP: *Opposite sense,* FAL: *False sense,* NO: *No sense,* IMP: *Imprecision,* TER/LEX: *Terminological/Lexical errors,* GRA: *Grammatical errors,* SPE: *Speed,* FILL/BK: *Overuse of pause fillers/Backtracking,* W: Weight value

count the various types of errors relevant to each category. As indicated in Table 1, there are three types errors for the category of *accuracy*. *Opposite sense* means rendering the opposite sense of a message in the source text. *False sense* indicates adding, omitting or substituting a message that makes sense in the target text but falsifies essential messages in the source text. *No sense* refers to making no sense at all in the target text. *Imprecision* means adding, omitting or substituting small details without undermining the source text essentials (Choi, 2005, p. 201).

The category of *expression* has two types of errors: terminological/lexical errors and grammatical errors. *Terminological errors* indicate the absence or inaccurate usage of terminologies used in specific fields whereas lexical errors relate to the inaccurate usage of common vocabulary outside the range of specialized terms. *Grammatical errors* relate to grammatical mistakes that occur in the linguistic structure of the target text (Choi, 2005, p. 202).

Under the category of *presentation*, only two types of errors, which are considered to be *measurable*, are taken into account for assessment purposes: *speed* and *overuse of pause fillers/backtracking. Speed error* is the time gap witnessed between the target text time and the source text time. Every additional 30 seconds in the target text is counted as an error. *Pause fillers* refer to the 'ums' and 'ahs' that are used during pauses or moments of hesitations. *Backtracking* means restarting or repeating the same word, clause or sentence. The number of instances in which *pause fillers* or *backtracking* occurs is counted as an error (Choi, 2005, p. 203–204).

As argued by Choi (2008), performance criteria should be given different weight values considering that *accuracy* should be regarded as the most determinant criteria in assessing performance. Therefore, as indicated in Choi's PAT (2005) above, a total weight value of 20 was used. Considering that 10 would be too small and 30 too big to divide the weight value into the subcategories, the weight value of 20 was considered to be adequate. The highest weight value of 10 is distributed to *accuracy*, followed by 6 for *expression* and 4 for *presentation.* Such different distribution of weight value is necessary in order to ensure that *accuracy* is not compromised by *expressions* or *speed.* The PAT again subdivides such weight value under each category. In terms of *accuracy,* out of the total weight value of 10, 6 is allocated to *opposite sense*, *false sense* and *no sense* all together, and 4 to *imprecision* considering that the former are significant errors than the latter. For the *expression* category, out of the weight value of 6, 3 is allocated to *terminological/lexical errors* and 3 to *grammatical errors*. Under the category of *presentation*, out of the weight value of 4, 2 is distributed to *speed* and 2 to *overuse of pause fillers/backtracking* (for more details, see Choi, 2005).

In the sample of Table 1 above, 5 errors that relate to *opposite sense*, *false sense* and *no sense* were detected. 5 errors correspond to the error range of 5

Table 1. Sample of the Performance Assessment Tool (PAT, adapted from Choi, 2005)

	ERROR COUNTS		RATING SCALE							W	COMPO-SITE SCORE x 0.1
ACCURACY											
OPP	0	5	Points	6	7	8(o)		9	10	6	8.4
FAL	5		Rating (ER)	Inferior (over 7)	Poor (6~7)	Acceptable (5)		Good (3~4)	Excellent (0~2)		
NO	0										
IMP	5		Points	6	7	8		9(o)	10	4	
	Rating (ER)	Inferior (over 9)		Poor (7~9)	Acceptable (6)		Good (3~5)	Excel-lent (0~2)			
EXPRESSIONS											
TER /LEX	2 Rating (ER)		Points	6	7	8(o)		9	10	3	5.1
			Inferior (over 3)	Poor (3)	Acceptable (2)	Good (1)	Excel-lent (0)				
GRA	1 Rating (ER)		Points	6	7	8		9(o)	10	3	
			Inferior (over 3)	Poor (3)	Acceptable (2)	Good (1)	Excel-lent (0)				
PRESENTATION											
SPE	2		Points	6	7		8(o)	9	10(o)	2	2.8
	Rating (ER)	Inferior (over 3)		Poor (3)	Acceptable (2)	Good (1)	Excel-lent (0)				
FILL/BK	22		Points	6(o)	7		8	9	10	2	
	Rating (ER)	Inferior (over 13)		Poor (10~13)	Acceptable (9)	Good (5~8)	Excel-lent (0~4)				
OVERALL PERFORMANCE SCORE											16.3

Test: midterm test; Student name: B; Source text time assessed: 2min. 30sec.(entire source text time : 3min. 2sec.); Target text time: 3min. 15sec.

and to the rating of 'Acceptable'. As a result, the student gained 8 points. 5 errors were identified for *imprecision*, which corresponds to the error range of 3–5 and to the rating of 'Good'. In this way, the student gained 9 points. The points were then multiplied by each weight value. For instance, 8 points, which was obtained from the subcategory of *opposite sense*, *false sense* and *no sense*, was multiplied by the weight value of 6, which resulted in 48. For the subcategory of *imprecision*, 9 points was multiplied by the weight value of 4, which resulted in 36. The two sub-weighted values, 48 and 36 were added up together, which resulted in 84. The composite score generated from each category was divided by a factor of 10 in order to fix the final overall performance score at 20.

The same calculation method applies to *expression* and *presentation*. For instance, as demonstrated in the Table above, 2 errors were detected for *speed* considering that the target text time surpassed 45 seconds compared to the original text. 2 errors in *speed* correspond to the rating of 'Acceptable' or 8 points. 8 points was multiplied by the weight value of 2, which resulted in 16 points. 22 instances were identified for *pause fillers/backtracking*, which correspond to the rating of 'Inferior' or 6 points. 6 was multiplied by 2 and resulted in 12 points. 12 points added up to 16 points made 28, which is finally divided by a factor of 10. In this way, the scores of each category are added together at the end to generate the overall performance score.

However, given that all we can observe is the result of the interpreting process, it is true that it may be hard to pinpoint whether the error simply results from the first category, a lack of *accuracy*, or the second category, a lack of *expression*. At times, the cause of the error could be both. As Choi (2006) points out, several or multiple elements, which are often inter-related, could be the cause of poor performance in interpreting (p. 277). For this reason, an inter-judge reliability test between the two raters was carried out to ensure that the results of the two raters were valid and reliable from each other on the whole.

Therefore, after the raters had individually finished assessing the student performances, their performance assessment results were subject to an inter-judge reliability test. The following are the results:

Table 2. Inter-judge reliability index of overall performance

Initial test	**Midterm test**	**Final test**	**Mean of combined scores**
≒ 0.76	≒ 0.69	≒ 0.81	≒ 0.75

2.6 Statistical analysis

A Repeated Measures of Analysis of Variance (Anova) test, which is a statistical tool used to analyze the differences among three or more means especially when the measurements from the same group are taken at different periods (Colman, 2003, p. 630), was implemented in order to analyze the evolution of performances between the three periods – initial, midterm and final tests. The p-value was set at 0.1 for several reasons. To understand the p-value, one must understand the null hypothesis, which assumes that nothing happened and that the results were obtained by mere chance (Crawley, 2005). As explained by Crawley, the p-value is a measure of how much the null hypothesis is credible. Therefore, "the smaller the p-value, the stronger is the evidence against the null hypothesis" (Kirkwood & Sterne, 2003, p. 62). The p-values that are usually used are 0.001, 0.01, 0.05 and 0.1. A p-value below 0.001 would provide a strong evidence against the null hypothesis whereas a p-value above 0.1 would provide a weak evidence against the null hypothesis (Kirkwood & Sterne, 2003, p. 75). The most frequently used p-value is 0.05. However, the choice of a p-value could depend on previous related studies and the research objective. In general, if it is recognized that the research concerned is a topic that has often been covered by a number of previous studies, a p-value smaller than 0.05 is likely to be chosen to support the research hypothesis. However, if it is judged that the topic has been rarely proved in the field concerned, a p-value higher than 0.05 could be chosen. Moreover, if the research objective is to simply observe a trend, a higher p-value can also be justified.

Against this backdrop, the p-value used in this experiment is 0.1 considering that little study concerning text length has been carried out in the field of interpreting and that the research objective is to simply observe the possible trends of interpretation performance when the length of the original text becomes longer and, thus infer whether text length could affect the performance of novice learners.

According to the Repeated Measures of Analysis of Variance (Anova) test results, no significant results were witnessed between the initial test and midterm, between the midterm and final test, and between the initial test and final test in terms of *accuracy*, *presentation* and overall performance scores. One noticeable fact, however, was witnessed in the category of *expression* scores. Following is the Repeated Measures of Anova test results in terms of the *expression* category.

Table 3. Repeated Measures of Anova test results

Analysis variable : expressions						
j	Obs	N	Mean	Standard deviation	Minimum	Maximum
Initial	11	11	5.591	0.252	5.100	5.850
Midterm	11	11	5.223	0.401	4.650	5.850
Final	11	11	5.332	0.357	4.800	5.850

The mixed procedure		
Class level information		
Class	Levels	Values
No	11	A B C D E F G H I J K
j	3	Midterm, final, initial

Type 3 Tests of fixed effects				
	Num	Den		
Effect	DF	DF	F value	Pr > F
j	2	20	3.09	**0.0675***

Differences of least square means									
Standard									
Effect	j	_j	Estimate	Error	DF	T value	Pr > |t|	Adjustment	Adj P
j	mid	fin	-0.1091	0.1521	20	-0.72	0.4814	Bonferroni	1.0000
j	mid	ini	-0.3682	0.1521	20	-2.42	0.0251	Bonferroni	**0.0753***
j	fin	ini	-0.2591	0.1521	20	-1.70	0.1039	Bonferroni	0.3116

As indicated above, significant differences existed only in the *expression* category (p-value=0.0675<0.1). To be more precise, significant differences were witnessed between the initial and midterm expression scores (Bonferroni multiple comparison p-value=0.0753<0.1). Considering that the mean of the expression scores between the initial and midterm tests dropped from 5.591 to 5.223, the results lead us to believe that *expression* scores of all test subjects witnessed a significant decline in the midterm test.

As described before in Section 1.1, performance can be affected not only by text difficulty but also by other various elements outside the text which can be

explained by the physical situation in which the text has been uttered and the interpreter's own ability. However, the physical situation played little role during the experiment since the test subjects were all exposed to the same physical situation in the classroom when the source text was uttered. The test subject's individual ability was not considered a major factor either provided that the test subjects were all novice learners in their first year of training with the same language combination. In terms of the difficulty of the text content, the texts used during the experiment were not too complicated or technical considering that all test subjects were novice learners in consecutive interpreting. This was in line with Seleskovitch's and Lederer's (1995) argument that extremely difficult topics or stylistically sophisticated texts should be avoided until students have reached a certain level and have mastered their interpreting techniques (p. 54) at the later stage in consecutive interpreting.

In this context, one distinctive factor that could possibly explain the significant decline in the midterm's *expression* scores is that the entire length of the source text uttered in the mid-term test measured approximately 50 seconds longer than the two other source texts used in the initial and final tests provided that the speakers spoke within the range of regular *speed*. This allows us to conclude that *expression* can be undermined when the text becomes longer. That is, students may put less attention to reproducing the message into the target language as they would do in shorter source texts and concentrate more on comprehending the source text message, when the length of the source text becomes longer. Although the statistical results show that when the length of source text became longer in the midterm test, all expression scores significantly declined, the *expression* scores did not show any significant progress between the midterm and final test despite the fact that the text used in the final test was shorter than the text used in the midterm test. Still, the mean of the *expression* scores slightly improved from 5.223 in the midterm to 5.332 in the final test, indicating that the trend of improvement was there. Such results imply that text length could possibly affect the performance of novice learners.

3. Research limitations and future studies

We should keep in mind that the decline in scores does not always mean a decline in learning progress. As Choi puts it,

Decline in performance scores does not necessarily mean that a student has not made progress in learning. That is, a student can still make progress even if performance scores decline. This usually occurs when variables other than the

student's interpreting ability that are outside his or her control, affect performance (Choi, 2008, p. 146–147).

We should also be mindful that, as Seleskovitch and Lederer (1995) pointed out, there are no clear-cut guides on what should be the proper length of texts given to students in class since the proper length depend more on the difficulty of the text subject or the progress made by students. In addition, the text length would not matter much once students reach a certain level and master their note-taking techniques (Seleskovitch & Lederer, 1995, p. 53). It is true that longer texts may even turn out to be easier than certain shorter ones that may be too concise or not long enough to grasp the speaker's message intent. Though longer texts cannot always be considered to be more difficult than shorter ones, provided that other variables are controlled, the mid-term test results demonstrate that we cannot exclude the possibility that text length could affect student performance since longer texts are likely to require more endurance for novice learners to process a speech than shorter ones, which also supports the research hypothesis indicated in Section 2.1.

The objective of this research has also been attained. We have observed the possible trends of interpretation performance when the length of the text became longer, and can infer that text length is likely to affect the performance of novice learners in consecutive interpreting.

Nevertheless, there are several research limitations inherent in this study. First, the size of the sample may not be considered to be sufficient. With more samples, this research could have obtained more significant results to support its hypothesis. However, as previously argued, the difficulty of obtaining samples in interpreting is not an easy task to address given the challenges in gaining access to reliable and valid data in this field. Second, it should also be noted that the test subjects were working from their B language, English, into their mother tongue, Korean. If test subjects were working from and to a different language, accuracy, presentation or the overall performance scores could also have been affected, thus leading to different results. Such possibilities should be continued to be examined in future studies.

On the whole, the research methodology used in this research and its subsequent results could lead us to suggestions on how researchers could assess the impact of text length and other multiple variables on interpreting performance, which may hopefully become a modest step towards providing guidance as to how future interpreters should be trained and how teachers should understand student performance when they are faced with different source text variables such as text length.

References

Altman, J. (1990). What helps effective communication? Some interpreters' views. *The Interpreters' Newsletter, 3,* 23–32.

Angelelli, C.V., & Jacobson, H.E. (Eds.). (2009). *Testing and assessment in translation and interpreting studies.* John Benjamin Publishing Company.

Choi, J-Y. (2005). Proposing a performance assessment tool for consecutive interpretation. *Conference Interpretation and Translation, 7(2),* 195–215.

Choi, J-Y. (2006). Metacognitive evaluation method in consecutive interpretation for novice learners. *META, 51–2,* 273–283.

Choi, J-Y. (2008). *The effect of the metacognitive grid on the learning curve for consecutive interpreting: A metacognitive approach to learning and evaluating student performance.* Unpublished Ph.D. thesis, Graduate School of Translation and Interpretation (ETI), University of Geneva, Switzerland.

Colman, A.M. (2003). Oxford Dictionary of Psychology. Oxford/New York: Oxford University Press.

Crawley, M.J. (2005). Statistics: An introduction using R. Chichester: John Wiely & Sons, Ltd.

Kalina, S. (2002). Quality in interpreting and its prerequisites: A framework for a comprehensive view. In G. Garzone & M. Viezzi (Eds.), *Interpreting in the 21st century: Challenges and opportunities* (pp. 121–130). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Kirkwood, B.R., Sterne, J. A. C. (2003). Essential medical statistics 2nd edition. Blackwell publishing company: Oxford.

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

Nord, C. (1991). Text analysis in translation: Theory, methodology, and didactic application of a model for translation-oriented text analysis. Amsterdam: Editions Rodopi B.V.

Seleskovitch, D., & Lederer, M. (1995). A systematic approach to teaching interpretation. The Registry of Interpreters for the Deaf.

Shelesinger, M., & Pöchhacker, F. (Eds.). (2011). *Interpreting, 13:1, Special Issue, Aptitude For Interpreting.* John Benjamin Publishing Company.

Translation versus Language Errors in Translation Evaluation

Tomás Conde[1]
University of the Basque Country

Within Translation Studies, scholars have usually dealt with diverse typologies of errors; most of them distinguish between translation and language errors. It is common to identify the former with phenomena of meaning, but some authors include in this group all errors that are not strictly related to language. This chapter compares the number of translation versus language phenomena (errors and good decisions) marked, and investigates if either of them has more impact on the overall quality judgment attached to each translation. It also explores whether group differences exist regarding the two types of errors included in the research. The results show that language errors are much more common in translation evaluation; however, translation errors contribute more on the variability of the marks issued. Translation professionals and potential addressees seem to base their marks only on translation issues, whereas translation teachers and students take both types of phenomena into account. Moreover, language errors play a key role on teachers' evaluations, whereas translation errors are more decisive among students. Finally, these effects are discussed and further research is suggested so as to confirm the main findings, which might be of interest both for professional and educational contexts.

Key words: Translation assessment, translation errors, language errors, analytical evaluation, empirical approach.

1. Introduction

Translation evaluation results in a judgment, whether numeric or not, on the quality of the translated text. As for its process, even though there are other tools – as well as mixed approaches that include the two most common types of instruments (Hajmohammadi, 2005, p. 219; Adab, 2000, p. 223; inter alia) – translation evaluation is usually carried out with the aid of either holistic or analytical methodologies.

1 tomas.conde@ehu.es

Holistic evaluation instruments classify each translation into any of the predefined levels within a scale, for example after revising its general characteristics. Analytical evaluation instruments, on the other hand, are based on the number of errors (and, sometimes, good decisions) that are first described, then quantified, and finally subtracted from the totality[2].

Thus, out of the two instruments the analytical one is more clearly based on the concept of error. Much has been said about translation errors, which are often categorized according to importance and nature. The former typologies depend on the importance attached by evaluators to errors, but a discussion of them is beyond the scope of this chapter (see, nevertheless, Conde, 2009, pp. 123–126 and Conde, 2011, pp. 70–71).

According to other typologies, Waddington (2001, pp. 311–312) claims that error nature is precisely among the recurring themes of translation evaluation studies; in particular, he refers to the following issues: 1) translation error types, 2) the relative nature of errors, 3) the need to assess the pragmatic level, and 4) the distinction between language and translation errors. It should be noted that there is no agreement on the latter two categories[3]; however, Pym (1991) and other researchers e.g. Vivanco et al., (1990), Larose (1998), House (1981), Kussmaul (1995) (all cited in Waddington, 2001) distinguish between language errors and translation errors.

Language errors are detected just by reading the target text; they are often equivalent to the errors about target language expression and consist of mistakes on vocabulary, syntax, grammar, punctuation, coherence, style, etc.

Translation errors are, however, explained by the existence of a previous text: the source text upon which the target text depends. They are usually, but not exclusively, what other authors (for example, Bastin, 2000, p. 234) call errors of meaning.

As previously mentioned, the distinction between translation versus meaning errors has attracted widespread attention but, perhaps, it would be more convenient to consider two terms that explain these cases more accurately: on the one hand, errors that are solely attributable to translation; on the other, errors that are common to other forms of written communication.

In order to issue their quality judgments, evaluators tend first to go through both the source and the target texts. This double contrast has become the most

2 Examples of holistic instruments are described by De Rooze (2003, p. 54) or Muzii (2006, p. 23); examples of analytical evaluation, by Beeby (2000, p. 189) or Choi (2006, pp. 278–279).

3 For instance, Nord (1991) discusses pragmatic, cultural and linguistic errors; Cruces (2001, pp. 817–821) deals with errors of sense and formal mistakes; Pym (1991, p. 281) distinguishes between binary and non-binary errors, etc.

common way to control not only those aspects that might be evaluated in non-translated texts, but also (and, maybe, particularly) those resulting from the fact that the corresponding text is a translation. Brunette, Gagnon, and Hine (2005, p. 31) describe a type of monolingual revision based only on expression issues, and also a type of bilingual revision, where original and translated texts are compared and, as a consequence, comprehension problems may be inferred.

Even though now and again it is possible to know that something has not been translated correctly (for example, when there are certain ambiguities, false friends or nonsenses), what normally happens is that the evaluator needs the source text to check the accuracy of the translation. But this is not always the case, basically because this double reading implies much effort, not to speak of the extra time that the evaluators often lack. Besides, translating is essentially a question of language; both concepts (language and translation) overlap so taking language errors as if they were alien to the activity of translating does not seem to make any sense. According to Williams (1989):

> […] an error of form can at the same time be an error of meaning and that a language error can cause a mistranslation or at the very least impede the reader's understanding of the translation […] (p. 25).

In any instance, two quality indicators could be distinguished in translation evaluation: those motivated by the fact that the text is a translation, and all the rest. Then the question that arises and gives meaning to this chapter is: Are opinions and marks conditioned by the incidence of phenomena restricted to translations or, instead, of the phenomena that may be present in other types of texts? One could argue that, when assessing, evaluators should focus on those skills that can be measured only in translated texts. In fact, if the questions that are common to all texts play a greater role in the marks awarded to translations, then it would be necessary to rethink the status of translation as an example of special communication, or alternatively the implicit criteria that make it possible to build an idea of the quality of the translated text.

This descriptive-relational chapter is part of a continuous body of research which aims to observe the process and result of translation evaluation performed by diverse populations. The results of the main research – which is fully described on a PhD thesis submitted at the University of Granada (Conde, 2009) – were that 1) subjects' behaviour and quality judgments show general tendencies, regardless of their population group; 2) the four groups also have special behavioural features; and 3) the length and the serialization of the task generate order effects both on the evaluators' behaviour and their quality judgments.

Some particular studies have followed the doctoral thesis; they deal with some aspects related to evaluation, as for example: order effects in the process and

result of lenient and demanding evaluators (Muñoz & Conde, 2006), the differences observed in the evaluation of texts on various subject matters (Conde, 2010), the impact of some features on the surface of translated texts (Conde, 2011), the relationship between quality and certain quantitative parameters (Conde, 2012) and the general behaviour of lenient and demanding translations evaluators (Conde, forthcoming).

These studies address some interesting questions regarding translation evaluation; the present chapter culminates in some manner the research conducted so far, aiming to offer a more accurate perspective of such a terra incognita: the impact of the strictly translatological issues on translation evaluation.

In the following section (Section 2), materials and methods are described; then (Section 3), results are presented and discussed; finally (Section 4), conclusions are presented.

2. Materials and methods

Data were extracted from a corpus comprising evaluations carried out on a total of 48 translations, grouped into 4 sets (regarding 4 English original texts) of 12 (Spanish) translations each. Two sets consisted of political texts for a wide readership; the other two were on technical communication, specifically on industrial painting techniques. Moreover, target texts had been translated by 4th year students in two courses ("Divulgative and Literary Translation English-Spanish" and "Scientific and Technical Translation English-Spanish"). Each text had to be translated within an hour; the teacher had suggested that students translated the texts as if it were for entering translation agencies or companies.

Once the assessment task was ready (not only the four translation sets, but also a final questionnaire that allowed for the contrast among quantitative results, biodata and personal opinions), it was sent to four groups of evaluators, all of whom were related, directly or indirectly, to this activity: translation students (25), professional translators (13), translation teachers (10) and potential addressees of the texts (40). The hardest stage of the investigation was probably searching for voluntary evaluators, as the disparate number of collaborators in each group suggests. Although the whole sample of subjects does allow for statistical generalizations of the data, the partial (per group) results, however interesting they may seem (for they point to general tendencies), should be viewed with caution[4].

4 Not all translations had been evaluated by all the evaluators (potential addressees worked only on half of the task, and some evaluators left the task after working on the first set); besides, not all translations had the same number of words (for they belonged to 4 different originals). Thus, as in Conde (2009, pp. 275–276), the results were multiplied by a compensation index.

Subjects were invited to "assess / correct / proofread / edit / revise [the texts] according to their beliefs and intuition, and to the best of their knowledge." The only requisites were that they had to:

1) process the translations in the order they had been given;
2) work on a whole set in a single session; and
3) classify the translations according to their quality into four categories: *very bad, bad, good* and *very good*.

These marks were afterwards converted into numerical values (1, 2, 3 and 4, respectively) to facilitate the statistical treatment of the data. A Microsoft Access database was designed and, for most analyses, extra files were created for the Statistical Package for the Social Sciences (SPSS, version 15.0).

The evaluation process was described in accordance with several parameters[5], most of which depended on two inter-related basic concepts. On the one hand, *action* was operationally defined as "any mark introduced by the evaluator in the text or file". On the other, *phenomenon* was defined as "what motivates or may motivate an evaluator to act on a particular text segment". Phenomena are usually errors, but evaluators act also to indicate good translation decisions or to enter other types of information (comments on the task itself, appeals to the researcher, etc).

The general analysis incorporated many parameters (*comments*, *reaction*, *saliency* and *scope*, among others), but only some of them are needed to test the hypotheses in question, which are the following:

- During the evaluation process, do subjects detect more phenomena that are exclusive to translated texts than those that are common to other text types?
- Regardless of the total amount, does any type of phenomena have a greater impact on the quality judgments issued?
- And finally, does the population group have any influence on the results?

Nonetheless, first *phenomena* must be grouped: those exclusive to written translation (from now on, exclusive) and those common to other types of written communication (from now on, common). These categories were built upon one of the most interesting parameters included in the main research: the nature, which consists of the interpretation given (according to the researcher) to the reason why the evaluator has worked onto a particular text segment[6]. Furthermore, a close examination of the evaluations made it possible to identify the following

5 See Appendix for an explanation of the parameters and categories included in the main research.

6 For more information on the main research, see the previous section and, especially, the author's doctoral dissertation (Conde, 2009).

categories (that could be grouped on the basis of multiple approaches[7]): *typos*, *punctuation, format, spelling, proper nouns, terminology, concordance, cohesion, syntax, weights* and *measures, appropriateness, clarity, usage, divergent interpretations, omission, perspective, unknown* and *combined phenomena*.

The first group (exclusive) would include phenomena that had been previously categorized as omissions (missing segments), divergent interpretations of the original and perspectives (disagreements in the translation decisions taken).

The second group (common) covers all other categories, with the exception of combined and unknown phenomena, for they cannot be correctly grouped: the former refer to several types of phenomena at the same time, and are usually introduced to summarize the pros (and cons) of the text; meanwhile, the latter only occurs occasionally, when an evaluator marks a text but the researcher is unable to interpret the reason for such a mark.

3. Results and discussion

This section provides an overview of the number of common and exclusive phenomena in the sample; afterwards, it contrasts the average of phenomena in each category with the global quality judgment; and, finally, it discusses the differences among the four population groups.

3.1 Common vs. exclusive phenomena

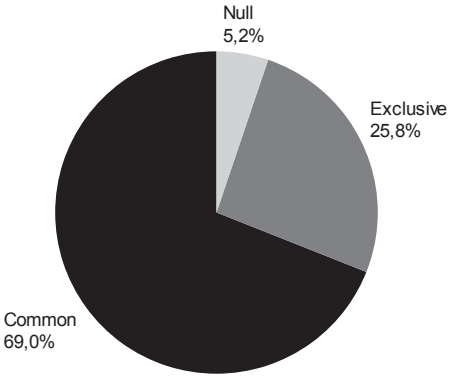
First of all, an analysis was conducted to determine which type of phenomena was more frequently marked by detailed evaluators[8], who amount to 45,839 total actions. Figure 1 shows the percentages of the different kinds of phenomena.

About 69 % of the phenomena could be found in other types of texts, too. Why is this the case? Detailed, analytical evaluators seem to mark all aspects that call their attention when revising translations, probably because of an excess of professional zeal. This may point to an overlap between the evaluator's functions and

7 See, for instance, Conde (forthcoming), for a distinction between normalised phenomena and non-normalised phenomena.

8 Evaluators completed the task with varying degrees of thoroughness. Only 55 subjects (*detailed* evaluators) carried out an explicit analysis of the texts before issuing the quality judgments. The rest (*concise* evaluators) marked the texts, but hardly worked on them. As in previous works (Conde, forthcoming), the count of phenomena and their contrast with quality judgments were based on detailed evaluators, for only they provide information on both the result and the process of their evaluation.

Figure 1. Phenomena per type



those of the reviser, for these concepts are related (Lee, 2006, p. 411; Brunette et al., 2005, p. 30, inter alia).

On the other hand, just one out of twenty phenomena are null. Also among detailed evaluators (Conde, 2009, pp 341–342), most of them were combined actions that the subjects performed to summarize each translation's features and introduce their quality judgments. As significant differences are expected neither within texts (as all of them had to be equally qualified) nor within groups (as all had to issue quality judgments), the following analyses focus on the two main categories: exclusive and common phenomena.

3.2 Impact on the quality judgment

Linear regression models were used to determine the correlation and significance of the explanatory variables (exclusive and common phenomena) with the dependent variable (quality judgment). Figure 2 shows the plots of both types of phenomena against the quality judgment: it seems that the greater number of phenomena, the lower the qualification obtained, especially in the first case.

A regression was performed to investigate the associations between the average quality judgments and the average number of phenomena identified, distinguishing between common and exclusive phenomena. Following a stepwise multiple regression procedure, the coefficient obtained equals 0.724 ($F[2,45] = 59.000$, $<p .05$). However, one of the two variables (exclusive phenomena) incorporated explains over the 65 % of the variance (Table 1).

Figure 2. Exclusive and common phenomena compared to quality judgment with regression

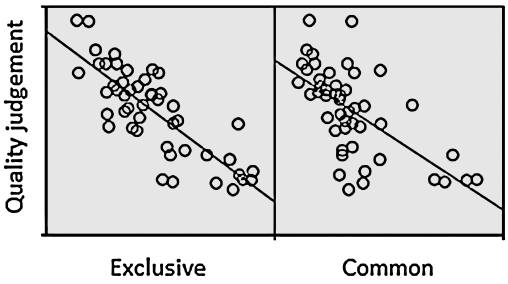


Table 1. Exclusive vs. common phenomena contribution

	R2	T	Sig.
Exclusive	0.656	-7.291	0.000
Common	0.724	-3.338	0.002

Therefore, data confirm the trend pointed by Figure 2, but the most informative factor in identifying quality judgments is the number of exclusive phenomena.

The question arises – when contrasting this result with that in Section 3.1- on how such a high percentage of phenomena marked in the text have such a low influence on the quality judgments; and vice versa how it is possible less than a quarter of what is corrected has such a significant effect on quality judgments.

Perhaps, it is clear to many evaluators that correcting and evaluating are different operations and, even though they use the assessment process to correct everything that lies in their path, when they have to mark texts, they focus especially on those aspects which are unique to translations.

In any instance, evaluators seem to implicitly differentiate errors on the basis of their worth, which agrees with what some authors (Fernández, 2005, p. 41; Seong, Lee, & Lee, 2001, p. 13) say about the relationship between error importance and its nature[9].

9 For example, major errors are often related to deficiencies in the knowledge of the working languages, whereas others consider those related to text contents seriously. The latter idea appears to be more common in Academia, where certain studies suggest that both translation students and translation teachers consider more serious errors those concerning literalness and the omission of relevant information.

3.3 Population group

With respect to the total percentage of each type's phenomena in the four population groups, no significant differences were found (Table 2).

Table 2. Exclusive and common per population group

Population groups	Common		Exclusive		Total
	No.	%	No.	%	
Students	20,078	73.6	7,213	26.4	27,291
Addressees	1,401	60.3	924	39.7	2,325
Translators	5,165	75.1	1,715	24.9	6,880
Teachers	4,978	71.5	1,984	28.5	6,962

If anything, there is a higher percentage of exclusive phenomena in addressees, which could be due simply to a matter of quantity, as this group carried out fewer actions than the other groups (see the last column), that is, more subjects were required in this group to reach the same percentage.

Separate linear regressions were modelled for each group (Table 3) to establish the relative contribution of variables.

Table 3. Different linear regression models for the four population groups

Population groups	Variables	R2	T	Sig.
Students	Exclusive	0.611	-6.668	0.000
	Common	0.686	-3.283	0.002
Addressees	Exclusive	0.306	-4.505	0.000
Translators	Exclusive	0.216	-3.560	0.001
Teachers	Common	0.607	-6.879	0.000
	Exclusive	0.798	-6.512	0.000

Only one predicting variable was found in two of the four groups:

- Potential addressees: exclusive phenomena, the analysis of variance (ANOVA) confirming the statistical significance of the findings ($F[1,46] = 20.291, < p$

0.05); on the other hand, the goodness of fit, or its R2 hardly explains the variance.

- Professional translators: exclusive phenomena ($F[1,46] = 12.671$, $p < 0.05$); the R2 is very bad.

For the two other groups there are two predicting variables:

- Students: $F(2,45) = 49.269$, $p < 0.5$. The most important variable is exclusive phenomena, which alone explains over 61 % of the variance.
- Teachers: $F(2,45) = 45.408$, $p < 0.5$ explaining almost 67 % of the total variance. In isolation, the other variable (common phenomena) is now the one showing a greater impact on the variance.

Regarding translation professionals and potential addressees, the R2 determination coefficient, which is a statistical measure of how well the regression line approximates the real data points, has only limited usefulness as a measure of the impact of the explanatory variables (below 0.4, the model fit is considered bad or very bad). Students and teachers show a good R2 value (above 0.5). Further, both variables are incorporated into the regression model in these two groups.

The differences in the four population groups are striking. On the one hand, neither potential addressees nor professional translators seem to have too much into account the exclusive phenomena (and hardly consider the other phenomena) to mark text quality. They may be aware that, instead of merely "correcting" the text, they are expected to get a sense of its quality. They somehow perform a more superficial evaluation, quite similar to that carried out within professional translation environments, where a bad translation is not that without errors, but the translation where "the total errors are within the desired threshold in a quality index" (Muzii, 2006, p. 24). Moreover, addressees and translators happen to include concise (holistic) evaluators. Although results were calculated only among those using analytical (non-holistic) systems, their approach seems less thorough than that chosen by the other two groups.

Besides, whereas the most relevant phenomena for students are those exclusive to translation, teachers seem to have common phenomena more into account. These subjects, who could feel also that their reputation was at stake, might want to pass on the care in the expression to their students, who do not identify it yet so much with the concept of translation quality. Another possible explanation for this is that teachers, realizing the low level of students, felt more responsible for their feedback, which is not limited to the source/target text relationship.

4. Conclusions

During the evaluation process, certain types of errors and good decisions (that is, of phenomena) can appear only in translated texts. This chapter has shown that the two types of phenomena analyzed do have different relevance and incidence rates for evaluators.

The phenomena that may be present also in other types of texts are much more common than those that are to be found only in translations. Nevertheless, linear regression analyses showed that the quality judgments issued by the evaluators were better explained by the so-called exclusive phenomena. Future research should investigate this rather paradoxical finding in more depth, for it seems that, even though evaluators are aware of the difference between revision and evaluation, they cannot help editing the text as they read it in order to qualify it.

This chapter has also shed some light on the four population groups' peculiarities, although the corresponding results should be handled with caution, for not all the groups had a statistically significant number of subjects. Therefore, the following interpretations are strictly personal and are aimed only to encourage reflection.

There seems to be a first effect related to environment: for extra-academic evaluators (professional translators and potential addressees), common phenomena do not have a direct impact on quality judgment, and the impact registered for the other phenomena is small. This could be due to a more superficial approach towards the evaluation task (none of them are concerned about learning the craft, but about the quality of the text as such). Or else, evaluators may think that they are not expected to assess what is not exclusive to translation.

On the other hand, the quality judgments issued by academic evaluators (both students and teachers) are influenced by both types of phenomena. Nevertheless, there is an essential difference between them: the largest contribution for teachers' marks is obtained by common phenomena. There are two probable reasons: 1) students, who are not yet aware of the importance of linguistic, grammatical or stylistic factors, focus on what they have to learn immediately (i.e., translating), and 2) teachers do their best to mark those aspects, in order to make their hypothetical students aware of the importance of delivering, not only an accurate translation, but also a clean, precise text.

As in previous studies (Conde, forthcoming; Conde, 2012; Conde, 2011), the present chapter falls within the scope of a broader range of activities that are not explicitly aimed to look at the impact of exclusive and common phenomena on quality judgments. Future initiatives should correct this aspect and could include other interesting points that here, due to space constraints in this chapter, have

not been addressed; for example, it would be interesting to check whether the impact of the two types of phenomena on quality judgments varies depending on the translated text type. Several genres or texts on different subject matter could be used. Or else, specialized texts and texts for a widespread reader might be compared.

This being so, evaluation is still a much-debated topic within Translation Studies: it is present in each and every environment where translation is carried out, and there are still many questions unanswered on this controversial but fascinating activity.

References

Adab, B. (2000). Evaluating translation competence. In C. Schäffner & B. Adab (Eds.) *Developing translation competence* (pp. 215–228). Amsterdam & Philadelphia: John Benjamins.

Bastin, G. (2000). Evaluating beginners' re-expression and creativity: A positive approach. *The Translator, 6*(2), 231–245.

Beeby, A. (2000). Evaluating the development of translation competence. In C. Schäffner & B. Adab (Eds.) *Developing translation competence* (pp. 215–228). Amsterdam: John Benjamins.

Brunette, L., Gagnon, C. & Hine, J. (2005). The GREVIS Project: Revise or court calamity. *Across Languages and Cultures, 6*(1), 29–45.

Choi, J. Y. (2006). Metacognitive evaluation method in consecutive interpretation for novice learners. *Meta, 51*(2), 273–283.

Conde, T. (2009). *Proceso y resultado de la evaluación de traducciones.* Unpublished doctoral dissertation, Universidad de Granada.

Conde, T. (2010). Tacit technique on the evaluation of technical texts. In L. Gea, I. García Izquierdo & M. J. Esteve (Eds.) *Linguistic and translation studies in scientific communication* (pp. 197–217). Oxford: Peter Lang.

Conde, T. (2011). Translation evaluation on the surface of texts: A preliminary analysis. *JosTrans, 15,* 69–86.

Conde, T. (2012). Quality and quantity in translation evaluation: A starting point. *Across Languages and Cultures, 13*(1), 67–80.

Conde, T. (forthcoming). The Good Guys and the Bad Guys: The Behavior of Lenient and Demanding Translation Evaluators. To appear in *Meta*.

Cruces, S. (2001). El origen de los errores de traducción. In D. Pujante González, E. Real Ramos, D. Jiménez Plaza & A. Cortijo Talavera (Eds.) Écrire, traduire et représenter la fête (pp. 813–822). Valencia: Universitat de València.

De Rooze, B. (2003). *La traducción, contra reloj.* Unpublished doctoral dissertation, Universidad de Granada.
Fernández Sánchez, M. M. (2005). La traducción especializada "bajo sospecha": Valoraciones negativas de un grupo de expertos. *Confluências, 2,* 28–45.
Hajmohammadi, A. (2005). Translation evaluation in a news agency. *Perspectives: Studies in Translatology, 13*(3), 215–224.
Lee, H. (2006). Révision: définitions et parameters. *Meta, 51*(2), 410–419.
Muñoz Martín, R. & Conde Ruano, J. T. (2006). Effects of serial translation evaluation. In Peter A. Schmitt & Heike E. Jüngst (Eds.) *Translationsqualität* (pp. 428–444). Frankfurt: Peter Lang.
Muzii, L. (2006). Quality assessment and economic sustainability of translation. *Rivista internazionale di tecnica della traduzione, 9,* 15–38.
Nord, C. (1991). *Translation as a purposeful activity: Functionalist approaches explained.* Manchester: St Jerome.
Pym, A. (1991). Translation error analysis and the interface with language teaching. In C. Dollerup & A. Loddegaard (Eds.) *Teaching Translation and Interpreting Training Talent and Experience* (pp. 279–288). Amsterdam: John Benjamins.
Seong, C., Lee, S. & Lee, H. (2001). *A survey of translation classes and evaluation criteria at GSIT Korea.* Retrieved from http://isg.urv.es/cttt/cttt/research.html.
Waddington, C. (2001). Different methods of evaluating student translations: The question of validity. *Meta, 46*(2), 311–325.
Williams, M. (1989). The Assessment of Professional Translation Quality: Creating Credibility out of Chaos. *TTR: traduction, terminologie, rédaction, 2*(2), 13–33.

Appendix: Phenomena and category of the main research

With regard to *actions*, the following variables were considered:

- The *number of actions* introduced by evaluators. Actions could be either *changes* or *comments*.
- Types of changes: text modifications, which could be grouped as:
  - *Feedback-oriented:* information for the reader; categories:
    - *Linkups:* marks used just to introduce comments.
    - *Highlights:* the evaluator merely identifies segments, without providing any information about why they call the attention to those segments.

    - *Classifications:* the evaluator marks segments in a systematic way, aided by a colour or format code.
  - *Product-oriented:* direct modifications of the body text:
    - *Additions:* the evaluator introduces words or sentences that are aimed to improve the translation.
    - *Suppressions:* the evaluator deletes words or sentences.
    - *Substitutions:* the evaluator deletes words or sentences and introduces others instead.
- Comments, which introduced information that did not become direct part of the text. There are five parameters regarding comments:
  - *Number of comments.*
  - *Location,* that is where comments were recorded:
    - *At the beginning:* before or just at the beginning of the body text.
    - *At the end:* after or just at the end of the body text.
    - *In the margin:* around the body text.
    - *In the text:* in the body text.
    - *Separately:* in a separate page, email or at the end of the digital file.
  - *Contribution,* depending on their function; there were two groups of categories:
    - *Investigation* comments (which depended on the specific context of the task and its instructions): *communications* to the researcher, *explanations* on the evaluation system and general quality *assessments.*
    - *Revision* comments (specific but indirect improvements of textual fragments): *alternatives, corrections* and *solutions*.
  - *Source,* or on whom the comment depends:
    - *External:* authority, linguistic convention, norm or logic shared by the community.
    - *Personal:* whenever the evaluator did not mention external sources.
  - *Certitude:*
    - *Certain:* when the comment was not uncertain.
    - *Uncertain:* when the comment expressed hesitation, uncertainty, insecurity or irony on the part of the evaluator.

Other parameters were connected with *phenomena:*

- *Scope,* depending on the length of the text portion affected by the phenomena; categories:
  - *Sentence,* or shorter segment within a sentence.
  - *Paragraph:* two or more sentences in a row.
  - *Text:* the whole translation.
  - *Set:* the twelve translations in a set.

- *Task:* all translations evaluated.

- *Nature;* categories where grouped as:
  - *Normalized* (there is an authority, normally the Spanish Royal Academy, who sanctions a proper option):
    - *Cohesion,* regularity and method of the solutions.
    - *Concordance,* grammar coherence and correlation.
    - *Format:* boldface, underlining, blanks, bullets, etc.
    - *Proper nouns,* spelling or syntax of proper nouns.
    - *Punctuation.*
    - *Spelling.*
    - *Syntax,* combinations of words and structures.
    - *Terminology,* specialized terms.
    - *Typos,* misprints.
    - *Weights and measures,* cases about writing numbers either in figures or in words.
  - *Non-normalized,* categories:
    - *Appropriateness,* register, tone or personal taste.
    - *Clarity,* incomprehensibility or confusion.
    - *Divergent interpretations* of the original.
    - *Omission,* lack of information.
    - *Perspective,* transfer decisions implying different scopes.
    - *Usage,* loan translations, collocations and rhetoric preferences.
  - *Others* (they could not be categorized neither as normalized nor as non-normalized phenomena):
    - *Combined* phenomena: several causes could be ascribed to the same action.
    - *Unknown,* unclear phenomena.

- Evaluator's *reaction* to the phenomena detected:
  - *Negative:* correcting, criticizing or marking a segment.
  - *Positive:* indicating that a segment has been correctly translated.
  - *Very negative:* insisting on the bad quality of a segment.
  - *Neutral:* communicating thoughts or ideas without criticizing or praising the rendering.

- *Saliency:* the biggest or smallest coincidence in the subjects at marking or detecting phenomena; categories:
  - *Zero* level of coincidence, or singular phenomena.
  - *Very low* level of coincidence (marked by two or three evaluators).
  - *Low level* of coincidence (between 4 and 12 evaluators).
  - *Medium level* of coincidence (between 13 and 25 evaluators).
  - *High level* of coincidence (between 26 and 40 evaluators).

Crosscutting parameters:

- *Quality judgment,* with its four categories: *very bad, bad, good* and *very good*.
- *Order* effects, analyzed at two levels:
  - *Sets: DP1, CT2, DP3* and *CT4*.
  - *Stretches,* or subdivisions within the sets: *I* (with the translations 1 to 4 of each set; *II,* 5–8 and *III,* 9–12).
- *Segmentation,* or divisions within the texts:
  - *Sections*:
    - *Initial:* first third, including titles and footnotes.
    - *Central:* text between the initial and final sections.
    - *Final:* last third, including footnotes, closing salutations and truncated sentences.
  - *Poles:* the first *(title)* and last *(ending)* segments of the translations.
  - *Typography,* or text segments according to their typographical treatment: *outstanding* (highlighted) or *regular*.

Part II
Applications of Assessing Language Translation and Interpreting in Local Systems

The System of Authorizing Translators in Finland

Leena Salmi[1]
School of Languages and Translation Studies, University of Turku, Finland

Ari Penttilä[2]
Authorized Translators' Examination Board, Helsinki, Finland

The system of authorizing translators in Finland has undergone several changes during the 45 years of its existence. It has evolved from a translation test measuring language skills rather than translation competence to an examination containing translation assignments as well as a test on professional practices. The current Act and Decree entered into force in 2008. The system is supervised by the Authorized Translators' Examination Board, operating in conjunction with the Finnish National Board of Education. This chapter examines the features of the current system of authorizing translators in Finland. The authorized translator's examination consists of two translation assignments and a test on the candidates' knowledge of the authorized translator's professional practices. The translations are done on a computer; printed and electronic dictionaries, other reference material and the Internet may be used, but the use of machine translation, translation memories and personal contacts is not allowed. Knowledge of professional practices is tested with multiple-choice questions, without access to the Internet. The tests are prepared and evaluated by qualified evaluators authorized by the Finnish National Board of Education.

Key words: authorization of translators, translation quality assessment, assessment scales, professional translating, legal translation, translators' professional practices.

1. Introduction

The systems of authorizing or certifying translators vary between countries. Brief descriptions of the practices in some European countries are collected in a document compiled by the European Union of Associations of Translation Companies (EUATC, 2009). In this chapter, we describe the system currently in force in Finland. It was revised recently (2008) in order to better correspond to the work of authorized translators.

2. Assessment of translations

The assessment of translations, or Translation Quality assessment (TQA), is a practice conducted both in professional working life and during translator training.

1 leena.salmi@utu.fi
2 ari.penttila@as-english.fi

It is also a field of inquiry within translation studies, with different approaches and a considerable amount of research. Hatim and Mason (1997, pp. 199–200), for instance, make several distinctions between different types of assessment, serving different purposes. They distinguish between formative and summative assessment (the first being continuous and the second providing information for decision-making); between proficiency testing (for testing performance) and achievement testing (for testing knowledge related to a certain curriculum); and between norm-based and criterion-referenced assessment. Williams (2001) describes existing TQA models and divides them into two main types: models with a quantitative dimension, and non-quantitative, textological models, according to whether or not they incorporate quantitative measurement. Lauscher (2000) makes a distinction between equivalence-based and functional approaches.

In fact, as House (1997, p. 1) points out, any assessment of translations reflects the evaluator's view of translation – what translating is and what the purpose of the translation is. This is also what Lauscher's (2000) distinction is based on. In equivalence-based assessment, the purpose of translating is seen as the reproduction of the source text "as closely as possible" (Lauscher, 2000, p. 151). In functional assessment, translating is seen as the rewriting of the source text according to a certain function for a defined target audience. Function is defined "with regards to the use of the target text in the target culture situation" (ibid, p. 156). Equivalence-based TQA models, in Lauscher's (2000) classification, include models such as the ones put forward by Katharina Reiß (1971) and Juliane House (1997), while functional models follow the functional or skopos-theoretic approaches developed by scholars such as Holz-Mänttäri (1984), Reiß and Vermeer (1984) and Nord (1991). We will not go into detail in describing all the models cited, but discuss some of them that we find relevant to the context of this chapter, the assessment of authorized translation and the system in use in Finland. In this chapter, we will use the terms 'evaluation' / 'evaluator' and 'assessment' synonymously.

Among the quantitative models described by Williams (2001) there are two models in use in Canada in somewhat similar contexts to those in Finland. The first one is the model developed by the Canadian government's Translation Bureau, the Canadian Language Quality Measurement System (Sical). The Bureau provides translation, interpretation and terminology services principally to the Canadian Parliament, but also to the private sector (Translation Bureau 2009). The Sical system is used both as an examination tool and as an internal tool for the Bureau to assess the translations it delivers. The Sical has "a scheme based on a twofold distinction between (1) transfer and language errors and (2) major and minor errors and on the quantification of errors" applied to samples of text (400 words per text) (Williams, 2001, p. 330). Williams, formerly in charge of running the Sical system (Mossop, 2004, p. 190), describes the changes made to Sical in the 1980's and 1990's from a

maximum number of 12 errors in an acceptable translation to "zero defects" (Williams 2001, p. 331), i.e. delivering translations that contain no errors. According to Williams, the Sical system is still in use "for examinations as well as for predelivery and performance evaluation purposes" (*ibid*).

The second model, similar to the Sical, is the one used by the Council of Translators and Interpreters of Canada (CTIC) for its translator certification examinations. It differs from the Sical model in that "no single error may be considered sufficient to fail a candidate" (*ibid*).

Williams (2001, pp. 335–343) also presents a proposition for an argumentation-centred TQA model of his own. The assessment is based on argument macro-structure and rhetorical topology: the source and target texts are analysed and compared, and the assessment consists of defining whether these elements are rendered in the target text if they are present in the source text. The results of the comparison are presented in a table with two columns for marking the presence of the elements and a third column for an assessment. This yields "an assessment of overall quality" of the translation (Williams, 2001, p. 342). The model has later been developed with more testing (Williams, 2004) and to allow more flexibility in the elements assessed as well as a weighted scoring system (Williams, 2009).

2.1 Specific features of authorized translating

Authorized translating can be defined as an "activity that has as its result a legally valid translation meant to be used as a tool for decision-making within an authority or in a legal procedure" or "in international transfer of documents" (Hietanen, 2005, pp. 17, 74). The guidelines given on authorized translating by the Finnish Association of Translators and Interpreters refer to "legally valid translations" and state that the status of translated document must not change in the process of translating a legally valid document (SKTL, 2009). Heli Mäntyranta (2010), a translator authorized within the Finnish system(s) since 1976, reflects on the tasks of an authorized translator saying that there are two procedures involved. First of all, the text is translated in such a way that the barriers related to language and culture can be surpassed to enable the reader of the translation to understand what the original text is about. Second, the translator's signature certifies that the translated document is legally as valid as the original document, and can be used, in the target culture, to determine the rights and duties of the holder of the document (Mäntyranta, 2010).

As authorized translating can be seen to constitute a special field of translating in that there are certain rules to be followed as to the form of the translations (e.g., naming the translation and signing it) and in that the equivalence of the content of the translation with the original is important (the translations need

to be precise and contain "everything" without omitting anything), it could be thought that its assessment needs to be equivalence-based. However, it can also be seen that in authorized translating, the function is to produce a target text that reflects the source culture as closely as possible. Therefore, the same applies as was mentioned earlier, that it is the conception of the evaluator that counts.

3. The origins of the Finnish system3

The current Finnish Act on Authorized Translators entered into force on 1 January 2008. However, this was not the first Act of its kind in Finland. The practice of granting translators the right to work as 'sworn translators' goes back several decades. This right was originally granted by Local Register Offices – without any specific test or examination – to persons who were judged to be competent to work as officially recognized translators. Often these persons were language teachers.

The first legal provisions governing sworn translators were laid down when the Act on Sworn Translators was passed in 1967. By that time, specific colleges known as 'language institutes' had been established to provide translator training, and the Act on Sworn Translators stipulated that students who had graduated from a language institute and had earned the highest grade in the final examination would be entitled to apply for the sworn translator's right. A second way to obtain this right was to take a separate translation examination implemented by the Ministry of Education. In the end, only a handful of translators applied for the right on the basis of their language institute degree. Taking the examination was by far more popular.

The Act did not stipulate any formal requirements for candidates who wished to test their skills in the examination. Candidates were required to translate a one-page text, which was usually an excerpt from a newspaper or magazine article dealing with some general or topical subject. The translations were written by hand, and no dictionaries or any other reference materials were allowed during the examination.

By the early 1980s, it was generally recognized that the sworn translators' examination system needed revision. Thus, after some deliberation, a new Act was passed in 1988. Two notable changes were made in the translation test: Instead of one text, candidates had to translate two texts, and the use of dictionaries and printed reference material was allowed. One text to be translated was of a general nature, while the other was from a specific field that the candidate could

3 The origins of the system and the system itself (Section 5) have previously been discussed by Penttilä (2008). At that time, no examinations had yet been organized according to the new system.

choose from among four alternatives: law and administration; economics and commerce; medicine and biology; and technology and industry.

4. Legislative preparation for the latest reform

It did not take long before it was again considered that the system of authorizing translators had become outdated, in particular as modern IT technology and the use of the Internet had changed translators' working methods radically. The types of texts used in the examination were also criticized. Since the texts were often excerpts of articles from magazines and newspapers, it was claimed that they did not correspond to the kind of texts that authorized translators need to translate in real-life situations. These texts are often various certificates, register extracts, court decisions, contracts, etc.

Preparations for a new Act on authorized translators started in late 2004, when the Ministry of Education appointed a working group to survey the current situation and make proposals for the development of the system. The working group members represented a wide range of stakeholders, including universities, State administration, and professional translators working in the field.

The Act currently in force was drafted largely on the basis of this working group's recommendations, and on the basis of a second working group appointed in 2007 by the National Board of Education, which was to become the supervising authority in the new system. The Government Bill was submitted to the Finnish Parliament in June 2007, and the Act and Decree on Authorized Translators were passed in December 2007.

5. Current system of authorizing translators

The authorized translator system is supervised by the Authorized Translators' Examination Board, which operates in conjunction with the Finnish National Board of Education. The Examination Board

- organizes the authorized translator's examinations;
- confirms the examination results;
- grants the right to work as an authorized translator;
- monitors authorized translators' work;
- keeps a register of authorized translators; and
- decides on the form and the text of the stamp to be used by authorized translators (recommendation).

In order to be accepted as authorized translators, applicants must meet certain basic requirements: they must be of age and be of good repute; they must be permanent residents of Finland or one of the Member States of the European Union, or some other state of the European Economic Area; and they must have passed the authorized translator's examination.

Passing the authorized translator's examination is not required if the applicant has completed a Master's degree (or some other higher university degree) that includes at least 60 credits of translation studies (theoretical and practical university studies in translation).[4] Translator training programmes exist for the following languages (together with Finnish): English, French, German, Italian, Russian, Spanish and Swedish. These studies must include at least six credit points (equivalent to ECTS credits) in authorized translation. The right to work as an authorized translator can in this case be granted only for the language pair included in the degree and only in the applicant's working language A.

Authorization is granted for five years at a time. Before authorization is granted, the applicant must give an authorized translator's affirmation. Authorization can be continued on application for at most five years at a time. Continuation requires that the applicant has been working as an authorized translator and meets the other requirements for obtaining the authorized translator's right.

The authorized translator's examination demonstrates the applicant's command of the examination languages and other knowledge required in order to work as an authorized translator. The examination is open to anyone who is a permanent resident of Finland, a Member State of the European Union or another state of the European Economic Area. At present, the examination is arranged once a year. The examination is taken in one language pair at a time, in other words either from a domestic language (Finnish, Swedish or Sami) into a foreign language or from a foreign language into a domestic language. Taking the examination is subject to a charge (280 euros in 2011). The fee covers some of the costs for organizing the examination.

The segment of the examination measuring the other knowledge required in order to work as an authorized translator consists of 20 multiple-choice questions testing translators' professional practices. The questions pertain to the working procedures of authorized translators, the division of responsibility between clients and translators, ethical principles of authorized translation, legislation, methods of quality assurance, certification of translations, and developments in the field of translation.

The segment of the examination measuring language and translation skills consists of two translation tasks which are translated one at a time. One task con-

4 In Finland, translator training is provided at university level, with admission by entrance exam.

cerns the field of law and administration and the other one is selected between the special fields of economics, technology and medicine.

The text length per translation task is approximately 2,000 characters including spaces. The translation is done on a computer. The Internet and electronic and printed sources may be used, but the use of machine translation, translation memory systems, and personal contacts to other people via e-mail or other messaging systems is prohibited. This is controlled by supervisors moving around in the classroom during the test. If a candidate is found doing something not allowed, he or she is expelled from the classroom and failed. Each translation assignment must be completed within 2 hours and 45 minutes.

5.1 Examinations arranged to date

To date, four examinations have been arranged under the current system: one in each of the years from 2008 to 2011. In the first and third examination, candidates translated from foreign languages into domestic languages, whereas in the second and fourth examination, the direction was from domestic languages into foreign languages. An examination between two domestic languages can be taken every time when an examination is held.

Table 1. Authorized translators' examinations arranged to date

Year	**Candidates**	**Language pairs**	**Pass rate**
2008	85	13	23 (27 %)
2009	106	24	16 (15 %)
2010	99	18	8 (8 %)
2011	89	18	11 (12 %)

As can be seen from Table 1, the pass rates differ considerably between the years. The material does not yet allow any statistical analysis, and so the reasons for the variation can only be hypothesized. It has been suggested that since there had been a break of a couple of years between the last examination arranged according to the old system and the first examination in the new system, the candidates who took the first examination were those who were the most motivated and had already been waiting for this opportunity. This would then be one reason explaining the relatively high pass rate in the first examination. The fact that it was the

first examination of the new kind may also have discouraged people with less experience and confidence from participating in the examination.

The long-term average pass rate in the examinations arranged according to the old system for authorizing translators, in force between 1988 and 2007, was 21.7 per cent.

6. Evaluators

The examination segments are prepared by qualified evaluators appointed by the National Board of Education. Evaluators need to hold a Master's degree, have experience in translating documents in the languages in question, and participate in evaluator training. Exceptionally, a Bachelor's degree can be accepted; for example, in languages that are not widely taught in Finland, such as immigrant languages (Somali, Vietnamese, etc.). The evaluator training is organised by the National Board of Education in the form of seminars.

Evaluators are appointed to work within a specific language pair; for example, between Finnish and English. The appointment is granted for five years at a time.

The evaluators' tasks include the preparation and the assessment of the different parts of the examination. One of the evaluators of a certain language pair usually prepares the text to be translated from the foreign language into domestic languages, and all evaluators participate in the assessment. For the tasks from Finnish or Swedish into foreign languages, the same source text in the official language (one text in Finnish, one in Swedish) is used for all the translation tasks into the foreign language within the same speciality.

Evaluators are both university lecturers in translation studies or languages, and professional translators. Many of the translator trainers working as evaluators also teach the authorized translation courses in their universities, and taking part in both helps in keeping up-to-date on the matter.

7. Preparation of tasks for the examination

As mentioned above, the examination consists of a test of the candidates' knowledge of the authorized translator's professional practices (multiple-choice questions) and of two translation assignments. All three parts must be passed at the same time in order to pass the examination.

Both translation assignments are from an LSP domain. One text deals with the law and administration specialty and the other specialty is chosen by the candi-

date when registering from the following: 1) economics, 2) medicine, 3) technology.

- Texts in *law and administration* include court decisions, court case reports, contracts, legal documents, pre-trial investigation material, diplomas etc.
- Texts in *economics* include extracts from corporations' annual reports, financial reports, articles of association, financial contracts, documents about economics etc.
- Texts in *technology* include manuals, patent applications, reports in the fields of engineering and technology etc.
- Texts in *medicine* include medical case histories, medical reports, medical certificates, reports on medical research results etc.

The texts to be translated are as authentic as possible. Authentic texts cannot, obviously, be used as such, but names and other personal information must be deleted or made invisible. All texts must also be new, as all the texts used as translation assignments are published after each examination on the website of the National Board of Education. In order to have texts that are more or less equal in length, it has been decided that they should be approximately 2,000 characters in length (including spaces). This requirement often makes it difficult to use authentic tasks, as typical translation assignments are either shorter (e.g. diplomas) or longer (e.g. court orders). It is, however, possible to shorten a longer text, as well as to create authentic texts out of templates available (for example, a template of a marriage contract completed with fictional names and the length adjusted to approximately 2,000 characters).

8. Assessment

Each multiple-choice question has three choices, with only one correct answer in each question. The percentage of correct answers in order to pass is defined by the Authorized Translators' Examination Board separately for each examination. Since the questions vary from year to year, the percentage of correct answers required may also be adjusted slightly to reflect the difficulty of the test.

All translation assignments are evaluated by two evaluators. Optimally, one of them is a native speaker of the source language, and the other of the target language. If this is not possible (this has sometimes been the case with some languages not spoken widely in Finland), both evaluators need to have a good knowledge of the other language involved.

The translations are evaluated using a two-fold evaluation system where translations are assessed both for their content and the quality of the target language translation. In this respect, the system seems most closely to resemble the Sical model in use in Canada, (cited in Section 1), where a distinction is made between transfer and language errors (Williams, 2001, pp. 329–330). In terms of Hatim and Mason's (1997, pp. 199–200) classification, the system can be described as proficiency testing for the purpose of summative assessment.

For evaluating the content, the main focus is on the equivalence to the original: whether all parts of the text have been translated and whether all the elements expressed in the source text are rendered in the target text. As the original document needs to serve as a legally valid document in the target culture, equivalence is considered important.

The target language is assessed in terms of the acceptability of the target text when compared against the usage and norms of the target language. This applies especially to grammatical correctness. As the texts to be translated usually explicitly deal with or at least implicitly reflect legal systems that are different from the one in the target culture, the translations also must be consistent with the system of the source culture. Therefore, for example, names of institutions should not be rendered by the equivalent institution in the target culture. For example, rendering the French first degree court *Tribunal de Grande Instance* as the Finnish *Käräjäoikeus* or the English *Magistrates court* would be considered an error; a more adequate solution would be to leave the original name and complement it with an explanatory translation which refers to the level of the court but does not exist in the Finnish system: *Tribunal de Grande Instance alioikeus*.

Evaluators use a scale with a number of different error types that describe the categories of possible errors. The error types were originally drafted in the working group within the National Board of Education in 2008, using as a basis the system suggested by one of the group members, Andrew Chesterman (2001) in the journal of the Finnish Association of Translators and Interpreters. The categories are discussed in the evaluators' assessment seminars, held once a year, and adjusted when needed.

Currently, the content equivalence scale contains 8 different error types and the acceptability scale 7 (see Appendix). Error types differ in weight: some error types are considered more important than others, and all errors are weighted by the number of points associated with the error. The scale is from 1 to 9 (leaving out 7 and 8). 1-point errors are typos or defects in idiomatic expressions that do not affect the comprehensibility of the meaning. The 6-point and 9-point errors are errors in the interpretation of the meaning that may change the meaning of

the whole text (for example, an erroneous translation of the title of the document). They can lead to failing the candidate altogether.

In addition to marking the errors and the points, evaluators give a suggestion as to whether the translation is passed or failed. On the evaluation form for each candidate, the evaluators are also asked to indicate whether they consider the translation assignment to be easy, demanding or very demanding. When assessing this, the evaluators take into account the challenges posed by the subject matter, vocabulary and genre, as well as the time that the candidates had at their disposal (2 hours 45 minutes). There is no predefined maximum number of errors. The limit for acceptable points is decided by the Authorized Translators' Examination Board after all the translations have been assessed by the evaluators and the Board has gone through the comments on the evaluation forms. The translations must be more or less acceptable, without any severe errors – the texts the candidates produce should be acceptable as legally valid translations.

The translations are evaluated in their entirety, even if there are severe errors at the beginning of the text that would fail the candidate. It is thought that the candidates have the right to receive feedback about their examination, especially if they have failed.

All assessment takes place anonymously. The candidates write their translations on a computer, and the evaluators receive them on paper by mail. Each candidate is given a number to be used instead of their names, and the evaluators only see these numbers. When candidates receive feedback on their translations, the evaluators' names are not indicated.

When compared to the TQA problems that Williams (2001, pp. 328–329) lists, it can be said to the credit of the Finnish system that seems to respond to most of the criticism Williams makes. The whole text is evaluated (not just the samples); assessment on failing or passing is not made on the basis of the number of errors only; there are different levels of severity in the error assessment and multiple levels of assessments. The system is built for the purposes of the examination, and therefore it would probably not be suited for translation suppliers. To our knowledge, the system has not been tested in professional contexts. However, some of the evaluators responsible for the courses related to authorized translating at the universities use it in assessing the translations students produce.

9. Discussion

The principal goal of the system of authorizing translators is to ensure the availability of high-quality, legally valid translations in the official languages of the European Union, in languages spoken by immigrants living in Finland, and in the

principal languages outside Europe. At present, there are about 3,000 authorized translators registered in Finland. The great majority of them have obtained this right either by passing the old sworn translator's examination or the authorized translator's examination preceding the current system. It can be assumed that this number includes many translators who do not actively work as authorized translators or may have even given up translation altogether. Authorizations have been granted in nearly 90 language pairs, the most common being between Finnish, Swedish and English. Other common languages are Russian, German and French.

It is natural that authorizations concentrate in languages that are commonly studied at school and in which it is possible to obtain university-level translator training. To widen the range of languages among authorized translators, more varied language education would be needed and more sources (e.g. dictionaries) should be available, especially in languages spoken by immigrants in Finland. Passing the authorized translator's examination is demanding and requires solid knowledge of both source and target languages and cultures.

When the current system was introduced in 2008, challenges were expected at least in three areas:

- Drawing up the multiple-choice questions testing candidates' knowledge of the authorized translator's professional practices;
- Arranging the examination in a computer environment;
- Finding and training an adequate number of evaluators for selecting the translation assignments and for grading the translations.

After four examinations, it can be said that all of these issues have required effort, but none of the problems has been insurmountable. Technical issues have perhaps been the easiest to solve: the examinations have usually been held in the computer classes of universities or other educational institutes where enough technical assistance has been available. In contrast, meeting the demand for qualified evaluators has proved to be challenging in some language pairs not widely spoken in Finland. At present, the register of evaluators kept by the National Board of Education contains about 80 names. The register is constantly updated and supplemented both to widen the selection of languages available and to ensure that the responsibility for selecting the texts and grading the translations does not always rest with the same people.

Apart from the preparation of the multiple-choice questions, the selection of the actual texts for translation has also posed some challenges, since the texts need to meet a considerable number of criteria: the length of the text is 2,000 characters, including spaces; it should constitute a continuous whole; it should be as authentic as possible and represent the types of texts that authorized translators

encounter in real-life work; the text should have a sufficient number translation problems, i.e. it should not be too simple; and the texts used in different years should be comparable with each other, i.e. passing the examination cannot be easier in one year than in another year. Finding a text that meets all these criteria, or even most of them, can be a demanding task.

As regards the system of evaluating the translation assignments, it can be criticized for using what Hatim and Mason (1997, pp. 199) call "a 'points-off' system", which, according to them "bears only a very indirect relation to the test taker's ability to translate". Also, it can be criticized for assessing only products and not the process (such as the reference material and the information seeking methods used by the candidate), but in a test situation this would be somewhat difficult. However, one idea might be to ask the candidates to mention the sources they used for a certain number of (predefined) terms or phrases, and then assess the reliability of these sources.

The selection of texts also differs depending on whether the translation assignment is from domestic languages (Finnish, Swedish) into foreign languages or vice versa. In the former case, the same text in Finnish or Swedish is translated into all target languages represented in that particular examination, whereas in the latter case there are as many different texts in each category as there are source languages present. Both cases have implications in terms of equality. It can be argued that translating a Finnish text dealing with an administrative or legal topic into most other European languages, particularly into Swedish, is simpler than translating the same text into, say, Chinese, Persian or Somali. Difficulties arise from both cultural differences and from the lack of sources between the two languages concerned, or from the overall lack of written material available in some languages.

In the second case, where texts are translated from foreign languages into domestic languages, the issue of equal treatment stems from the great number of source texts. When there may be up to 20 different source texts, how can it be ensured that they are commensurable in terms of translation problems?

These issues of equality have been addressed in two ways: firstly, by selecting evaluators who have sufficient theoretical and practical experience in the translation sector; secondly, by holding training events for evaluators where they can meet their colleagues, can discuss any problems encountered and can agree on uniform procedures. A one-day seminar is also arranged each year in February for evaluators before they finalize their grading of the previous year's examination. This helps evaluators to adopt uniform practices in their grading work. These training events are organized by the National Board of Education.

10. Conclusion

In this chapter, we have described the system of authorizing translators currently in force in Finland, as revised in 2008. Authorization is now granted on the basis of an examination assessing both knowledge of working as an authorized translator and practical translation competence. The translation assignments are as authentic as possible. This change from a translation test of magazine and newspaper articles (Penttilä, 2008, p. 2) as well as the change to give the possibility to authorization on the basis of a university degree in translation studies have been welcomed by both practicing translators and translator trainers. Although criticism may be addressed to the system of assessing the translation assignments, the system is constantly being developed.

References

Chesterman, A. (2001). Polemiikkia virallisen kääntäjän kokeesta. [Discussion on the examination of official translator]. *Kääntäjä 9*, 10.

EUATC (2009). Practice in parts of Europe on sworn translations, notorisation and apostille. Document available on the Language Industry Web Platform of the European Commission. Retrieved from http://ec.europa.eu/translation/LID/index.cfm?fuseaction=main.PublicationDetail&PBL_ID=363

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

Hietanen, K. (2005). *Virallinen kääntäjän paljon vartijana. Ammattitoiminnan ja auktorisointimenettelyn yhteensovittamisen haaste*. [The Licenced Translator Safeguarding the Public Interest – The Challenge of Reconciling Professional Practice and Method of Authorisation] Doctoral dissertation, University of Tampere, School of Languages and Translation Studies. Retrieved from http://acta.uta.fi/pdf/951-44-6436-2.pdf

Holz-Mänttäri, J. (1984). *Translatorisches Handeln: Theorie und Methode*. Helsinki: Suomalainen tiedeakatemia.

House, J. (1997). *Translation Quality Assessment*. A Model Revised. Tübingen: Narr.

Lauscher, S. (2000). Translation Quality Assessment: Where Can Theory and Practice Meet? *The Translator 6*(2), 149–168.

Mossop, B. (2004). Compte rendu: Malcolm Williams: *Translation Quality Assessment: An Argumentation-Centred Approach*. *TTR: traduction, terminologie, rédaction, 17*(2), 185–190. Retrieved from http://id.erudit.org/iderudit/013278ar.

Mäntyranta, H. (2010). Kiinalaisia sopimuksia ja ghanalaisia adoptiopapereita. [Contracts from China and adoption documents from Ghana] *Kääntäjä* 3, 4.

Nord, C. (1991). *Text Analysis in Translation: Theory, Methodology, and Didactic Application of a Model for Translation-Oriented Text Analysis*. Amsterdam and Atlanta: Rodopi.

Penttilä, A. (2008). Seeking an optimal system for certifying translators. Finnish experiences over the past 40 years. *Proceedings of the XVIII FIT World Congress*. Publication on CD-ROM.

Reiß, K. (1971). *Möglichkeiten und Grenzen der Übersetzungskritik. Kategorien und Kriterien für eine sachgerechte Beurteilung von Übersetzungen*. Munich: Max Hueber Verlag.

Reiß, K. & Vermeer, H. J. (1984). *Grundlegung einer allgemeinen Translationstheorie*. Tübingen: Max Niemeyer. Translation into Finnish: Mitä kääntäminen on, 1986 by P. Roinila, Helsinki: Gaudeamus.

SKTL (2009). *Laillisesti pätevien käännösten laatimisohjeet*. [Guidelines for producing legally valid translations] Retrieved from the website of the Finnish Association of Translators and Interpreters: http://www.sktl.fi/@Bin/55747/Laillisesti+p%C3%A4tevien+k%C3%A4%C3%A4nn%C3%B6sten+laatimisohjeet.pdf

Translation Bureau (2009). About Us. Web page of the Translation Bureau / Bureau de la Traduction. Retrieved from http://www.btb.gc.ca/btb.php?lang=eng&cont=282

Williams, M. (2001). The Application of Argumentation Theory to Translation Quality Assessment. *Meta 46*(2), 326–344.

Williams, M. (2004). *Translation Quality Assessment: An Argumentation-Centred Approach*. Ottawa: University of Ottawa.

Williams, M. (2009). Translation Quality Assessment. *Mutatis Mutandis 2*(1), 3–23. Retrieved from http://aprendeenlinea.udea.edu.co/revistas/index.php/mutatismutandis/article/viewArticle/1825

Appendix

Grading system used by the evaluators of the National Board of Education for the Authorized Translator's Examination since 2011.

Error category	**Error type**		**Points**
Equivalence of content (C); precise and faultless use of special terminology.	• An idea is completely misinterpreted	C1	9 p. leads to a failed examination
	• A wrong term leading to the misinterpretation of the translation	C2	9 p. leads to a failed examination
	• The translation function is disregarded, leading to an inadequate result	C3	6 p. may lead to a failed examination
	• Unfounded alternative translation equivalents, i.e. the choice is left to the evaluator	C4	6 p. may lead to a failed examination
	• An omission or an addition essentially affecting the meaning of the text, e.g. a general and crucial abbreviation is not translated	C5	6–2 p. depending on the severity of the omission; 6 p. may lead to a failed examination
	• An individual word/term that is imprecise, unsuitable or irrelevant for the content or culture but does not necessarily lead to the misinterpretation of the translation • An omission or an addition not essentially affecting the meaning of the text	C6	6–2 p.
	• Misinterpreted structure	C7	6–2 p.
	• Incomplete or erroneous equivalents for the cultural and social context of the source language	C8	2 p.
Acceptability and readability of text (A). General acceptability and readability of text; usage according to orthographical, morphological and syntactic norms; register and style correspond to the text function and the intended use of the translation.	• A structural error that is likely to cause misinterpretation	A1	6–4 p. depending on the severity of the error; 6 p. may lead to a failed examination
	• Inconsistent terminology or style	A2	6–2 p.
	• A spelling mistake that affects the interpretation of the text section	A3	4–2 p.
	• Inadequate translation in terms of the information structure of the text	A4	2 p.
	• A structural error that does not cause misinterpretation	A5	2 p.
	• Individual style errors and unidiomatic expressions	A6	2–1 p.
	• A spelling mistake that does not affect the interpretation of the text section	A7	1 p.

Translation Competence and the Practices of Translation Quality Assessment in Turkey

Nilgun Dungan[1]
Izmir University of Economics, Turkey

Opinions regarding what constitutes a 'good' translation differ significantly. Until recently, the focus of research on assessment has been on identifying what is understood by concepts such as equivalence and developing objective sets of criteria to judge translation quality. However, recent developments in the industry and academia call for a different perspective. Translation industry has been developing at an amazing speed due to advances in technology and globalization, intensifying the need to spell out standards to ensure the quality of the service provided. One such standard, EN 15038, for instance, specifies competences a translator must have in order to be employed by the service provider. In the meantime, many institutions of higher education in Europe are currently involved in an intergovernmental European reform process, known as the Bologna Process, which emphasizes the role of higher education in providing students with skills and attributes needed in the workplace. Bologna qualifications frameworks, comprising statements of learning outcomes and competencies that students have to show before they could earn their degree, make it necessary for many universities to review their curricula, based on discipline-specific competencies. The present chapter attempts to throw light on the issue of translation competence in terms of industry and training.

Key words: translation competence, quality assessment, evaluation, training, industry.

1. Introduction

Researchers in any academic field attempt to define their object of study before proceeding with their endeavors, and the field of translation studies is no exception. However, as any scholar who has attempted to define translation as a process and a product could attest, such is not an easy undertaking. No universal definition of translation exists, and opinions regarding what constitutes a 'good' or 'quality' translation, as well as those about the central concepts of evaluation and its purposes, differ significantly. While there is some consensus on certain criteria of quality, such as clarity of ideas expressed and accuracy in terms of grammar and lexis, the field of translation quality assessment is quite problematic, because clarity and accuracy, too, are relative terms that resist definition and might take on various shades of meaning depending on the purpose and the context of translation. Depending on how translation is perceived, the significance attached to these notions has varied greatly throughout time (cf. Schäffner, 1998).

1 nilgun.dungan@ieu.edu.tr

Since the emergence of translation studies as an academic discipline, a great number of scholarly research has been conducted in 'pure' branch of Holmes' framework of translation studies. However, the 'applied' branch, especially translation criticism, including the evaluation of translations for different purposes (e.g. assessment of trainee translations, certification exams, reviews of published translations, and so forth) requires more critical and systematic research. The area of assessment offers challenges as well as new and exciting directions: there is no agreement on the central concepts of evaluation; many different approaches exist side by side, looking at translation from different angles, using different terminology, and treating translation either as a process or a product, all of which have serious implications for evaluation of translation quality. Many scholars still perceive it as a "probabilistic endeavor, one in which subjectivity constitutes the most salient criterion" (Arango-Keeth & Koby, 2003, p. 117).

Until recently, the focus of research on assessment has been on identifying what is understood by concepts such as translation equivalence and developing objective sets of criteria to judge translation quality. However, recent developments in the industry and academia call for a different perspective, particularly looking into the concept of translation competence and its implications on translator training and quality assessment.

2. Translation Competence

Derived from the Latin verb *competere*, meaning "to meet, agree," the term competence resists a universal, one-size-fits-all definition, much like translation and quality. Yet, competence is one of the latest truisms, a buzzword if you will, in industry and higher education these days: it is a must-have ability. A nice concept for sure, but how does one acquire it? How do we teach and assess a competence? How do we know when one has it or whether it even exists in practice? Current developments in the industry, international educational arena, and recent research in translation studies yield different interpretations of the term competence. In other words, while competence appears to be a catch-phrase for all involved, there is little consensus about its development, components or acquisition.

Nonetheless, the concept of translation competence is highly relevant in translator training and quality assessment. In the past two decades an increasing number of studies have been conducted regarding the definition, teaching and assessment of translation competence which currently drives many curricular decisions, particularly in Europe. Today in translation studies, translation competence is understood as "the set of knowledge, skills and attitudes that enable an individual to act as a professional translator, although there are scholars (e.g. Kiraly, 2000) who still

keep translation competence distinct from translator competence" (Palumbo, 2009, p. 31). Various studies (cf. Beeby et. al, 2000; Adab, 2000; Kiraly, 2000; Way, 2008; PACTE, 2011) have looked into the development and/or acquisition of translation competence. For instance, the second phase of the research carried out by PACTE Group (Process of Acquisition of Translation Competence and Evaluation) investigates "the process of acquisition of translation competence in trainee translators with the aim of developing a holistic model of the acquisition of translation competence" (PACTE, 2011, p. 3). PACTE Group defines translation competence as "the underlying system of knowledge required to translate" and considers it as

> the underlying knowledge system needed to translate and has four distinctive characteristics: (1) it is expert knowledge and not possessed by all bilinguals; (2) it is basically procedural knowledge (and not declarative); (3) it is made up of various interrelated sub-competencies; (4) the strategic component is very important, as it is in all procedural knowledge (ibid).

Another on-going project about translation competence is TransComp. Funded by the Austrian Science Fund and initiated by University of Graz in 2007, it is a longitudinal, process-oriented study which attempts to explore the development of translation competence and hopes to contribute to the "development of the methodology and model building in process-oriented translation studies by overcoming a number of shortcomings of previous studies" (TransComp, 2011, "The Project," par. 1). The project, which aims to use the insight into the components and development of translation competence for translation pedagogy and the improvement of curricula for translator training.

2.1 Translation industry, standards, and competence

Businesses supply goods or services, and in doing so, they bear certain ethical and legal obligations toward their customers and stakeholders. In terms of quality, these providers of goods and services need to meet global standards, i.e. they have to meet not only the needs and expectations of their customers but also certain international standards. For instance, many European companies require that their suppliers comply with standards set by the International Organization for Standardization (ISO), a nongovernmental entity based in Geneva, Switzerland. As providers of service, professional translation businesses are also expected to meet these obligations.

Due to advances in technology and globalization, translation industry has been developing at an amazing speed, intensifying the need to spell out standards to protect both the service provider and the clients of translations. ISO oversees a wide spectrum of standards, but one that concerns businesses most is the ISO 9000 family which concerns quality and customer satisfaction. For instance ISO 9001,

the standard for quality management systems, is the most established quality framework, and it is currently being used by around 897,000 organizations in 170 countries worldwide. Many translation businesses, which have been trying to establish a way to prove the quality of their service to their customers, have adopted ISO 9001.

In 2006 the European Committee for Standardization, the CEN, published EN 15038, a quality standard for translation services, which is now gaining worldwide acceptance. Translation companies which certify under the standard agree to audit by independent institutions to ensure the quality of the service that they provide. Under the heading "Human resources management," EN 15038 requires the translation service providers (TSPs) to "have a documented procedure in place for selecting people with the requisite skills and qualifications for translation projects" and to "ensure that the professional competences required by 3.2.2 are maintained and updated" (BSI, 2006, p. 7). Table 1 below lists the competences that translators, revisers and reviewers should have as specified by the standard in sections 3.2.2, 3.2.3, and 3.2.4 respectively.

The standard also requires that the stated competences "should be acquired through one or more of the following: formal higher education in translation (recognised degree); equivalent qualification in any other subject plus a minimum of two years of documented experience in translating; at least five years of documented professional experience in translating" (BSI, 2006).

The terms and definitions used in the standard have very interesting implications for the area of assessment and merit scholarly attention, especially in light of the recent developments in higher education. For one, in relatively larger translation companies, translators have multiple roles such as project manager, information specialist, terminologist, editor and quality manager in addition to that of a translator, each role requiring a different set of skills or competences. This means that "in order to prepare students for the actual working life in the translation industry, each of these competencies should be adressed during the course of their training" (Malmkjær, 2004, p. 32) at a time when developing "more efficient methods of translator training is a necessity which results from a shortening of degree programmes in translation as a consequence of the Bologna process" (TransComp, 2011, "The Project," para. 1).

The current market demands necessitate that translators assume different roles but also require a myriad of skills such as IT, management, and interpersonal skills in addition to many different competencies that need to be acquired during training. Malmkjær (2004) also argues that "[p]ractice in translation skills is not enough to make a professional. Professionals need to have a background in the history, theory, and methodology of the subject in order to give them insight into their role and thus to strengthen their self-image as professionals" (ibid, p. 33). As a result of the paradigm shift in education, significant changes have taken place in teaching methodologies and assessment procedures. Whereas the old paradigm was teacher-

centered, with focus on language, isolated skills and product, the new paradigm is learner-centered, with focus on communication, integrated skills, and process. Thus, researchers face the challenge of developing new methodologies to teach subjects such as history or theory in the new ways. Some even prefer to use the terms facilitator or educator instead of the word teacher in traditional, didactic approaches.

Table 1. Professional competences of translators as specified in EN 15038

3.2.2 Professional competences of translators
Translators shall have at least the following competences.
a) Translating competence: Translating competence comprises the ability to translate texts to the required level, i.e. in accordance with 5.4. It includes the ability to assess the problems of text comprehension and text production as well as the ability to render the target text in accordance with the client-TSP agreement (see 4.4) and to justify the results.
b) Linguistic and textual competence in the source language and the target language: Linguistic and textual competence includes the ability to understand the source language and mastery of the target language. Textual competence requires knowledge of text type conventions for as wide a range of standars-language and specialised texts as possible, and includes the ability to apply this knowledge when producing texts.
c) Research competence, information acquisition and processing: Research competence includes the ability to efficiently acquire the additional linguistic and specialised knowledge necessary to understand the source text and to produce the target text. Research competence also requires experience in the use of the information sources available.
d) Cultural competence: Cultural competence comprises the ability to make use of information on the locale, behavioural standards and value systems that characterise the source and target cultures.
e) Technical competence: Technical competence comprises the abilities and skills required for the professional preparation and production of translations. This includes the ability to operate technical resources as defined in 3.3. […]

3.2.3 Professional competences of revisers
Revisers shall have the competences as defined in 3.2.2 and should have translating experience in the domain under consideration.

3.2.4 Professional competence of reviewers
Reviewers shall be domain specialists in the target language.

Another market attempt to ensure that training meets the demands of the industry is European Master's in Translation (EMT) expert group, set up by the EU Directorate General for Translation (DGT) in April 2007. EMT expert group was initiated to respond to "stimulate the increase of quality translator education" and to encourage the cooperation of European universities offering programs in translation. EMT establishes a reference framework for the competences applied to the language professions and claims that "[the framework] sets out what is to

be achieved, acquired and mastered at the end of training" (EMT expert group, 2009, p. 3). The group claims to have "sought to be as explicit and clear as possible to prevent differences of interpretation (from trainers with different backgrounds, experiences and constraints), in order to facilitate the implementation of these competences and the evaluation of their application and to speed up the networking of programmes complying with the framework thus defined." EMT defines competence as follows:

> combination of aptitudes, knowledge, behaviour and knowhow necessary to carry out a given task under given conditions. This combination is recognised and legitimised by a responsible authority (institution, expert). The competences proposed in each of the six areas are interdependent. Thus, for example, the aptitude for taking reasoned decisions is horizontal; it applies equally to the provision of a translation service and to documentary research. They all lead to the qualification of experts in multilingual and multimedia communication. Together, they comprise the minimum requirement to which other specific competences may be added (for example in localisation, audiovisual translation or research) (ibid)

The competences specified by EMT to be included in designing learning objectives in training include translation service provision (interpersonal, production), language, intercultural, information mining, thematic, and technological (mastery of tools). (For a sample of how the relationship between course learning outcomes and program competencies based on EMT are stated, see Appendix I).

Finally, another project initiated by EU Lifelong learning Programme, called OPTIMALE (Optimising Professional Translator Training in a Multilingual Europe), is an Erasmus Academic Network involving 70 partners from 32 different European countries (including 27 within the EU). The project aims to monitor the changing nature of the translation professions in the age of the internet, social networks and increasing automation and act as a vehicle and stimulus for innovation and high quality in the training of professional translators. In order to tune the training objectives with the demands and expectations of the industry, an employer consultation questionnaire has been developed within the framework of the OPTIMALE project. In other words, consultation with stakeholders in the translation industry is seen as a means of optimizing undergraduate translator training programs.

2.2. Bologna Process and competence

Many institutions of higher education in Europe are currently involved in an intergovernmental European reform process, known as the Bologna Process. Initiated with the Bologna Declaration in 1999, the Process emphasizes the role of higher education in providing students with skills and attributes needed in the work-

place. Bologna qualifications frameworks, which comprise statements of desired learning outcomes and competencies that students have to show before they could earn their degree, makes it necessary for many universities review their curricula, based on discipline-specific competencies. The Process, which aims to create a European Higher Education Area (EHEA) based on "international cooperation and academic exchange," is a pledge to reform the structure of higher education systems in the member countries. Today, 47 countries are united under the Process which also involves European Commission, Council of Europe, UNESCO-CEPES (European Centre for Higher Education), representatives of higher education institutions, and quality assurance agencies.

The Process aims to facilitate mobility of students, graduates and faculty members, to prepare students for their future careers by increasing their employability, and to offer broad access to quality higher education. The reforms are about offering comparable degrees organized in a three-cycle structure (e.g. bachelor-master-doctorate), quality assurance in accordance with the Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG) and allowing transparency, i.e. recognition of foreign degrees and other higher education qualifications in accordance with the Council of Europe/UNESCO Recognition Convention.

The 1997 Lisbon Recognition Convention and pan-European transparency tools like the European Credit Transfer and Accumulation System (ECTS) and the Diploma Supplement (DS) play an important role in this context. The qualifications framework for the EHEA, and the Standards and Guidelines for Quality Assurance in the EHEA are important as well. Extended goals of the Process include undertaking work "in areas of broader societal relevance, such as the links between higher education, research and innovation; equitable participation and lifelong learning" (Bologna Process, 2010, "About the Bologna Process," para. 1). In other words, the social dimension of European higher education places an emphasis on equity and employability of graduates in a lifelong learning context.

Bologna action lines include the provision of qualifications frameworks, easily readable and comparable degrees in the three-cycle structure, mobility, quality assurance, employability, lifelong learning, and EHEA in a global context. Among these lines, however, employability deserves additional emphasis as it seems to point to a shift with regards to the philosophy of education. From the very beginning, one of the main goals to be achieved with the creation of the EHEA has been employability, and during the ministerial meeting in May 2007 in London, it was identified as one of the priorities for the period leading to the next ministerial conference in April 2009.

While employability can be defined in many different ways, for the purpose of the Process it is defined as "the ability to gain initial employment, to maintain

employment, and to be able to move around within the labour market" (Bologna Process, 2010, "Employability," para. 2). Within the context of the Bologna Process, the role of higher education then is "to equip students with skills and attributes (knowledge, attitudes and behaviours) that individuals need in the workplace and that employers require, and to ensure that people have the opportunities to maintain or renew those skills and attributes throughout their working lives. At the end of a course, students will thus have an in-depth knowledge of their subject as well as generic employability skills" (Bologna Process, 2010, "Employability," para. 3).

Higher education institutions and businesses would certainly both benefit from working together, and measures have been presented by the European Commission in 2009 to develop and strengthen co-operation between universities and businesses, as part of "wider efforts to support the modernisation of higher education" (http://ec.europa.eu). However, now it seems that the raison d'être of higher education is to ensure that the students graduate with knowledge and skills, or competences as termed in the Bologna Process, demanded by the labor market as the driving force, making employers, especially large corporations, and students the consumers of higher education.

Bologna qualifications frameworks, which comprise statements of desired learning outcomes and competencies that students have to show before they could earn their degree, have made it necessary for many universities to review their curricula — in terms of structures, programs and actual teaching — based on discipline-specific competencies identified by relevant departments within the context of a "tuning process". Learning outcomes are defined in the glossary of Tuning Educational Structures in Europe as "statements of what a learner is expected to know, understand and/or be able to demonstrate after completion of a process of learning […] learning outcomes must be accompanied by appropriate ASSESSMENT CRITERIA which can be used to judge that the expected learning outcomes have been achieved" (González & Wagenaar, 2003, p. 258).

The term competence is defined in the same glossary as follows: "[…] a dynamic combination of attributes—with respect to knowledge and its application, to attitudes and responsibilities—that describe the LEARNING OUTCOMES of an educational programme, or how learners are able to perform at the end of an educational process" (ibid. 254). The potential for learning outcomes and competences marks a shift towards a concentration on the learner itself rather than on teaching, traditionally based on transmission of knowledge. This paradigm shift has important implications for the institutions of higher education and academia, who should develop new ways of transferring not just an accumulated body of knowledge but also facilitating the acquisition of a variety of competences by students.

Key competencies, which combine both cognitive and non-cognitive components, are not explicitly stated in official documents, but they are closely related to enhancing employability of the graduates. One of the potential problems here is that so much attention has been given to the needs of the industry while identifying the specific competencies that it shifts the paradigm in terms of the role of education. In other words, the industry-oriented Bologna Process tends to move away from the traditional role assigned to higher education in attempting to prepare the students for life and transform society by effecting change. Another problem might be the applicability of a European model of education in countries, such as Malta, Spain and Turkey, which might not have the same educational philosophy and backdrop. Additionally, as Aboites (2010) puts it, the Process maintains the idea of the "single way of thinking, seen now in a single group of competencies that are considered valid for Europe [and other countries] without considering the enormous cultural, social and political diversity of the countries of those regions" (ibid, p. 443).

3. Institutional Practices: the case of Turkey

Turkey, which has been involved in EU education programs such as Lifelong Learning program and Erasmus, has been a full member of the Bologna Process since 2001, and several universities in the country are now a part of it, as required by The Turkish Council of Higher Education (CoHE), the main body responsible for higher education, chiefly on administrative and financial issues. Another national body is Interuniversity Council, coordinating the activities of universities and preparing regulations on education and research (Yağcı, 2010). Situated on the eastern part of the EHEA, Turkey has a large higher education system, composed of mainly universities. The current number of universities is 165- 103 state-funded and 62 non-profit foundation universities – bound by the same higher education law. As of 2011, the total number of students is 3,817,086 students (OSYM Statistics, 2011).

Since 2001, a series of reforms have been introduced in Turkey. In terms of the degree structure reform, which includes the introduction of three cycles and depends on the European Credit Transfer System (ECTS), Turkey did not face any major challenges meeting the structural requirements since the three cycle structure already existed in Turkey, with first cycle lasting for four years and the second for two years in addition to a short two-year cycle, offering an associate degree. On the other hand, ECTS, based on student workload and learning outcomes, is used in all universities parallel to the national credit system, based on contact hours alone, which is not fully compatible with the ECTS. This shows the

system is still in a transition period. Only four pilot universities have completed this transition with many others expected to follow suit. Many Bologna Process countries face difficulties due to the specific nature of their national educational systems; however, the highly centralized, top-down higher education structure in Turkey and the traditional approach to education held by many institutions of higher education appear to present even bigger challenges in the process of adaptation.

3.1 Translator training programs

In Turkey, over 20 universities currently offer undergraduate degree programs in translation and interpreting or translation studies, with approximately 3500 students enrolled in these programs. Given the sheer number of these departments, the need for trained faculty members is enormous. According to Turkish regulations, a graduate degree is required to teach in undergraduate programs, which precludes the full-time employment of experienced professionals from the field. However, since TS is a relatively new academic discipline, the number of institutions offering MA and PhD programs in the field is very limited. As a result, many of the faculty members in translation and interpreting programs come from a variety of academic backgrounds such as English Language Teaching (ELT), Linguistics, and Language and Literature Departments, which naturally affect the way they view, teach, and evaluate translations. In other words, depending on their background, method of teaching, and purpose of assessment, these instructors adopt various approaches to translation and have different perceptions of quality in translation, which, in turn, influence their notion of translation competence and their methods of assessing the quality of student translations. Conventional evaluation methods of faculty from different fields tend to focus on the concept of translation errors. However, opinions on errors vary considerably, and even among trainers, there seems to be very little consensus about what constitutes an "acceptable" translation, as confirmed by a survey designed to identify and compare the assessment practices of faculty members in translator training institutions in Turkey (Dungan, 2010).

The survey revealed that all the participants use certain criteria when marking student translations, but their criteria were based on completely different factors. For instance, those participants with a TS background took the entire text as the unit of translation, taking into consideration the skopos of the translation, text type and the shifts of expression in evaluating the translations. On the other hand, those with an ELT background displayed a tendency to evaluate student translations at the sentence level with sentence length, level of difficulty and structural complexity being the decisive factors, whereas those with a background in lin-

guistics tended to evaluate on the word and sentence level, too, but with an added emphasis on fluency, idiomatic use of language, and style. The differences in the criteria they used naturally resulted in strikingly different scores assigned to each target text in the survey. Another interesting point revealed by the study was that instructors with a TS background stated that they find such marking difficult since a considerable percentage of the grade they assign usually comes from the commentaries that students are required to write to accompany their translations. In the commentaries, their students are asked to explain their strategies and choices and to indicate the difficulties they face and how they overcome these. The absence of such commentaries, these instructors added, made it difficult to evaluate the translations effectively. On the other hand, the difficulty expressed by ELT and Linguistics background instructors was the level of subjective interpretation inherent in the assessment of the target texts. In other words, assigning a point value to each sentence on the basis of its complexity and deducting points for errors proved to be a rather subjective and complicated task. One instructor admitted that she was not pleased with the criteria she used but felt it was impossible to avoid the degree of subjectivity in deciding the value of each sentence and the number of points to take off for each error.

Orozco defines translation competence as the "underlying system of knowledge and skills needed to be able to translate" with the sub-competences being transfer competence, communicative competence in two languages, extra-linguistic competence, instrumental-professional competence, psycho-physiological competence, and strategic competence (ibid, p. 120, qtd in Arango-Keeth & Koby, 2003). However, as the results of the survey indicate, for more than half of the participants, translation competence is perceived simply as linguistic competence. The lack of consensus on the perception of translation quality as well as on the competences of a translator, the general tendency to evaluate the product of translation rather than the process, and the implications of the notion of quality instilled through summative assessment in training and its alignment with the needs and expectations of the translation industry are but a few reasons for further research in the field of quality assessment. While the findings of the survey are not conclusive by any means, they, nevertheless, point to the problematic state of translation quality assessment in Turkey, which appears likely to become even more compounded with the latest developments in higher education.

4. Conclusion

The quality standard EN 15038 was accepted by the Turkish Standards Institute (TSE) and went into effect as of October 12, 2006, endorsed by a number of

associations, including the Association of Translation Businesses (ÇİD). In the meantime, translation and interpreting departments at Turkish universities are currently attempting to redefine and outline key competencies expected of future translators, although traditionally translation competence has long been considered synonymous with linguistic competence.

Though not explicitly stated in official documents, these key competencies are closely related to enhancing employability of the graduates. In fact, so much attention has been given to the needs of the industry in identifying the specific competencies that the paradigm of the role of education is shifting, making the labor market the driving force as stated by Way (2008): "without a doubt, the main objective of translator training is to prepare our graduates to enter the professional market" (ibid, p. 89).

For the time being, competence, and competence-based training (CBT), is an integral part of curricular design, increasingly making institutions of higher education face the challenge of restructring programs, designing new curricula and syllabi and evaluating the progress of their students. As Way (2008) puts it "CBT will continue to provide an extremely important trajectory for future translator training research" (ibid, p. 90).

As European policy makers try to bridge the gap between the competencies required of graduates and the knowledge that university systems traditionally transfer to students so that higher education meets the needs of the market, another point that merits attention is the applicability of a European model of education in countries such as Turkey, which might not have the same educational philosophy and backdrop as other European countries. While higher education institutions and translation businesses would certainly benefit from working together, the paradigm shift and the one-size-fits-all approach to education without much regard to cultural, social and political diversity of such countries have important implications for both institutions of higher education and academia.

References

Aboites, H. (2010). Latin American universities and the Bologna process: from commercialisation to the tuning competencies project. In *Globalisation, societies and education*. 8(3), 443–455.

Adab, B. (2000). Evaluating translation competence. In C. Schäffner and B. Adab (Eds.), *Developing translation competence* (pp. 215–228). Amsterdam and Philadelphia: John Benjamins.

BSI (2006). EN 15038: 2006 *Translation services—Service requirements*. London: BSI (British Standards Institute).

Arango-Keeth, F. & Koby, G. S. (2003). Assessing assessment: translator training evaluation and the needs of industry quality assessment. In B. J. Baer (Ed.), *Beyond the Ivory Tower: Rethinking Translation Pedagogy* (pp. 117–134). Philadelphia: John Benjamins.

Beeby, A., Ensinger, D., & Presas, M. (2000). *Investigating translation*, Amsterdam: John Benjamins.

Campbell, S. & Hale, S. (2003). Translation and interpreting assessment in the context of educational measurement. In G. Anderman (Ed.), *Translation today: trends and perspectives* (pp. 205–224). Clevedon: Multilingual Matters.

Dungan, N. (2010). Evaluating the quality of trainee translators: toil and trouble. *Translation and interpretation in a multilingual context* (pp. 51–56). Bangkok: Chulalongkorn University.

European Master's in Translation (EMT) expert group. (2012, April, 30). Competences for professional translators in multilingual and multimedia communication. Retrieved January 15, 2012, from: http://ec.europa.eu/dgs/translation/programmes/emt/key_documents/emt_competences_translators_en.pdf>

Gile, D. (1995). *Basic concepts and models for interpreter and translator training*. Amsterdam/Philadelphia: John Benjamins.

González, J. and R. Wagenaar (Eds.) (2003) *Tuning Educational Structures in Europe. Final Report – Pilot Project Phase*. Groningen and Bilbao.

Kiraly, D. (2000). *A social constructivist approach to translation education*. Manchester: St Jerome.

Maier, C. (2000). Introduction. In C. Maier (Ed.), *The translator* 6(2), 137–148.

Malmkjær, K. (Ed.). (2004). *Translation in undergraduate degree programmes*. Philadelphia, USA: John Benjamins.

Orozco, M., & Albir, H. A. (2002). Measuring Translation Competence Acquisition. *Meta* XLVII(3), 375–402.

PACTE (2011). Results of the Validation of the PACTE Translation Competence Model: Translation Project and Dynamic Translation Index. In S. O'Brien (Ed.), *IATIS Yearbook 2010*. London: Continuum.

Palumbo, G. (2009). *Key terms in translation studies*. London: Continuum.

Schäffner, C. (ed) (1998). *Translation and quality.* Clevedon: Multilingual Matters.

TransComp: The development of translation competence. (2011, August 8). The project. Retrieved January 16, 2012, from: http://gams.uni-graz.at/tc

Yağcı, Y. (2010). A different view of the Bologna process: the case of Turkey. *European Journal of Education* 45(4), 588–600.

Way, C. (2008). Systematic Assessment of Translator Competence: In Search of Achilles' Heel. In J. Kearns (Ed.), *Translator and interpreter training: issues, methods, debates* (pp. 88–103). UK: Continuum.

Appendix I

The Relationship Between Course Learning Outcomes and Program Competencies

1. Being able to use advanced, field-specific theoretical and practical knowledge acquired,
2. Being able to analyze fieldspecific concepts and ideas through scientific methods and to interpret and assess data,
3. Being able to understand and use grammatical, lexical and semantic structures of the source and target languages,
4. Being able to define functions and meanings of social, geographical, historical and stylistic variations of the language
5. Being able to understand and analyze micro and macro structures, cohesion, coherence, social and cultural functions of various kinds of texts in source and target languages and to produce such texts
6. Being able to transfer theoretical knowledge and skills developed in different areas of expertise into the act of translation
7. Being able to understand texts in source language and to render these texts into target language by using the register appropriate for the meanings and functions of these texts
8. Being able to use current technologies such as translation memory systems, online resources, terminology banks, spell and grammar checks, internet, terminology database efficiently in all processes of translation and to follow the developments in this field
9. Being able to use knowledge and skills with regards to the social role of translator and the ability to follow job profiles and requirements of professional life
10. Being able to define stages and strategies about translation, to define problems in the translation process and to find solutions to such problems
11. Being able to make decisions, criticize and display creativity in translation process
12. Being able to access necessary resources in the translation process and to incorporate such resources into the translation process effectively
13. Being able to use a second foreign language at an intermediate level
14. Ensuring that students gain lifelong skills

Evaluating Assessment Practices at the MCI in Cyprus

Georgios Floros[1]
University of Cyprus

This chapter draws on the experience gathered by the implementation of the *Masters in Conference Interpreting* (MCI) at the University of Cyprus. As a starting remark, there seems to be some room for improvement as to the way the assessment procedures and criteria set by the European Masters in Conference Interpreting (EMCI) were applied in the MCI. Specifically, this chapter discusses the inadequate application of quality criteria such as *general background knowledge* and *booth manners*, as well as other problems relating to the general *rating scale* followed by the University of Cyprus. There is also some room for improvement as regards the speeches used for training and assessing, as quite some problems surfaced among others by the fact that the same speakers who delivered speeches during training were the ones who also delivered the examination speeches. To these ends, this chapter sets off by describing the structure, content and assessment procedures and criteria followed by the MCI, so as to highlight problematic areas. It then attempts to provide possible explanations and propose ways to deal with problems in view of a continuation of the programme in the future.

Key words: assessment procedures, criteria, assessment literacy, conference interpreting

1. Presentation of the Masters in Conference Interpreting in Cyprus

In 2003, the University of Cyprus responded to the call of European Union institutions (the European Commission and the European Parliament) to establish new interpreter training programmes around Europe by setting up and implementing such a programme between 2004 and 2007. Specifically, the initiative for this programme was taken by the European Commission's interpreting service and conference organizer (previously *SCIC* (Service Commun Interprétation-Conférences), today known as Directorate-General for Interpretation, or *DG Interpretation*) and the interpreting services of the European Parliament (today known as *Directorate-General for Interpretation and Conferences*) as a result of their efforts for closer collaboration with European Universities to train highly qualified interpreters in order to respond to the rising demands posed by the then recent enlargement of the European Union. The core idea was to train interpreters mainly for the new languages and member states of the European Union, by drawing on the expertise and organizational culture of Universities. For roughly the same reasons, a pilot project led by the above European institutions and some

1 gfloros@ucy.ac.cy

Universities offering high-quality interpreter training programmes had already been launched in 1997. This project later led to the formation of the EMCI (*European Masters in Conference Interpreting*) consortium. The EMCI Masters programme follows a specially designed core curriculum and a set of strict assessment procedures and assessment criteria to ensure quality. The consortium applies strict selection criteria to allow further Universities to participate. Notably, the programme designed and implemented at the University of Cyprus never applied to participate officially in the EMCI consortium, mainly due to organizational reasons. Nevertheless, the structure, content, assessment procedures and assessment criteria of the Cyprus programme largely reflected those agreed upon and followed by the consortium.

1.1 Overall design of the MCI

The *Masters in Conference Interpreting* at the University of Cyprus (MCI) was designed as a vocational postgraduate programme to be completed within one year (12 months), during which students of various backgrounds had to be trained in the two main modes of conference interpreting, i.e. consecutive and simultaneous. Each academic year would start in September of a calendar year and end in August of the next calendar year. The curriculum spread over three semesters (fall, spring and summer semester), in which the main focus was put on consecutive (fall and spring semester) and simultaneous (spring semester) interpreting practice, public speaking, theory of interpreting and a project on terminology (summer semester). Greek was offered as *A language*, English (and later Turkish) as *B languages*, while *C languages* varied over the three years, subject to availability of trainers and demand on the part of each year's candidates (Spanish, French, German and Italian).

An extremely important component of the programme – and at the same time perhaps its main strength – was the pedagogical assistance received by both supporting European institutions. In the framework of the pedagogical assistance scheme, highly experienced active interpreters working for the European institutions (as permanent staff or free-lancers) visited the University of Cyprus throughout the duration of the MCI for usually a half week's time to monitor students' progress, guide them with techniques and advice, assist the local trainers in their work and train speakers in delivering speeches tailored to the pedagogical needs of the programme. Another important aspect, to which particular attention was paid, was to have native speakers of all languages involved deliver the speeches. In addition, the European institutions granted to students of the MCI a one-week pedagogical visit to Brussels once

a year, where students were offered the chance to practise dummy-booth interpreting (i.e. without transmission of the output to the audience) in real conditions, under the guidance of in-house interpreters.

The assessment procedures of the above programme included an aptitude test (scheduled about one week before the start of each academic year), a mid-year (or interim) examination at the end of the fall semester (end of January), and a final examination at the end of the spring semester (June), with the possibility for a re-sit examination before the start of the new academic year. All tests and examinations had to be conducted by a committee consisting of the coordinators and teaching staff of the MCI and at least one representative from each interpreting service of the supporting European institutions, with test speeches being delivered by roughly the same group of speakers.

The aptitude test was designed as a recruitment test on the basis of a shortlist of candidates who fulfilled the formal criteria for application (recognized University degree and proof of knowledge for at least two C languages or one B and one C language). It was eliminatory and consisted mainly of a short oral examination, where candidates were asked to reproduce a short speech in their C, B and/or A languages into their A and/or B languages. Some additional questions testing their overall background knowledge were possible, as well as a short interview about their motives and general skills. Successful candidates would then enter the programme and start training. The mid-year examinations tested the students' progress in consecutive interpreting (note-taking and quality of presentation), which would give a safe indication of their ability to continue with simultaneous interpreting. These exams were not eliminatory, unless students themselves decided to drop the programme due to a weak performance, unpromising for a continuation. On the contrary, final examinations (both in the consecutive and simultaneous modes) were eliminatory, although a second chance was offered to borderline cases (re-sit).

1.2 Assessment criteria

As to the assessment criteria the MCI followed in all the above assessment stages, there seems to be some room for improvement. While the MCI committed itself to the criteria stipulated by the EMCI in its curriculum, the way these were concretely applied displays some inadequacies worth discussing. This chapter draws on the experience gathered over these three years of implementation of the MCI at the University of Cyprus and focuses on the evaluation of both the procedures and the assessment criteria followed. First, it will be argued that while quality criteria focusing on *coherence*, *accuracy* and *deliv-*

ery of the interpreting output were applied consistently, other issues, such as *general background knowledge* and *booth manners*, could not always be considered in an adequate way when assessing candidates or students. Another issue was posed by the rating scale used to assess the candidates' performance. Although it was necessary to apply the general rating scale prescribed by the University of Cyprus for all courses offered, the specificities and complexities of conference interpreting as a skill, as well as the level of performance required both by recruitment examinations of European services and by the free-lance market itself, make it rather difficult for such a rating scale to hold. It will be further argued that there is also some room for improvement as regards the type, quality and consistency of the speeches used for training and assessing future interpreters across the various language pairs offered, as quite some problems arose among others by the fact that the same speakers who delivered speeches during training were the ones who also delivered the examination speeches.

In the sections to follow, an evaluation of the assessment stages and assessment criteria will be undertaken with the aim to highlight problematic points. An attempt will subsequently be made to explain the results of this evaluation, as well as to suggest possible remedies in view of a continuation of the programme in the future.

2. Evaluation of assessment procedures and criteria

2.1 Statistical data and preliminary interpretation

Before starting an evaluation of assessment procedures followed by the MCI, it would be useful to gain an overview of some statistical data concerning the number of successful candidates (aptitude test) and students (exams) across the three years of implementation of the MCI, as in Table 1 below.

Table 1. Successful students per assessment stage

	Academic year 2004/5	*Academic year 2005/6*	*Academic year 2006/7*	
				Average
Aptitude test				
Candidates	10	12	8	
Admitted	5 (= 50 %)	4 (= 33.3 %)	5 (= 62.5 %)	**48.6 %**
Mid-year exams				
Students	5	4	5	
Successful	5 (= 100 %)	3 (= 75 %)	5 (= 100 %)	**91.7 %**
Final exams				
Students	4	3	5	
Successful	2 (= 50 %)	1 (= 33.3 %)	1 (= 20 %)	**34.4 %**
Re-sit exams				
Students	2	2	4	
Successful	2 (= 100 %)	2 (= 100 %)	3 (= 75 %)	**91.7 %**

Table 1 offers an analytical account of how candidates and students performed in the various assessment stages (overall four official assessment stages, including the aptitude test and the re-sit examinations). The following Table 2 provides a more concise picture of the correlation between the number of students admitted to each year's programme and the number of students who successfully completed it, both through the final examinations and the re-sit examinations.

Table 2. Rate of successful completion per academic year

	Academic year 2004/5	*Academic year 2005/6*	*Academic year 2006/7*
Number of students admitted	5	4	5
Number of students who completed successfully	4	3	4
Rate of successful completion	**80 %**	**75 %**	**80 %**

The two tables above at first sight reveal that a) the pool of candidates consisted of relatively good prospective students, since the average success rate in the aptitude tests was 48,6% (cf. Table 1), and b) training has been quite effective, since successful completion ranged between 75% and 80% (cf. Table 2). In other words, the data seem to be confirming Donovan (2003), who maintains that not every applicant can be turned into a good interpreter, however good the course. The data also seem to imply that the aptitude tests must have provided an "effective screening", at least since the rates of successful completion reveal that the aptitude tests have led to a relatively small number of "false positives, i.e. of successful candidates who should not have passed" (Monfort, Moraki, Pouttu, Wang, 2008, p. 20).

However, what the data cannot reveal is the number of "false negatives, i.e. of unsuccessful candidates who should have passed" (ibid.) the aptitude tests (cf. also Dodds, 1990). Nor can they reveal the number of "false positives" as concerns the final and re-sit examinations, i.e. the number of successful students who are rather incompetent to practice professional interpreting on the real market (cf. Clifford, 2005, p. 128). In fact, the three-year experience we gathered through training all these students allows us to assume that many of them did not manage to internalize the importance of so-called 'psychological factors' such as team-work and booth manners. Many of the students admitted to the programme showed an extreme weakness in managing stress, in collaborating with peers for self-study sessions and, generally, refused to accept that training in conference interpreting depends heavily on self-monitoring and self-study, beyond contact hours in the classroom (cf., for example, Bartłomiejczyk, 2007; EMCI, 2011). Furthermore, most of them had a rather hard time accepting critique and evaluation by the staff or their peers. The above was not only the case with the students who discontinued their studies immediately after the mid-year examinations in 2004 and 2005 (which explains the difference between the number of students who succeeded in the mid-year exams and the number of students who continued to the final exams in Table 1), but also with some students who completed the programme. The latter were perhaps the most 'perilous', as they sustained a rather unpleasant atmosphere, which, besides not promoting progress, did not contribute at all to their acquiring the necessary and highly important professional ethics.

To return to the 'false negatives' of the aptitude test, there is no intention of questioning the jury's decisions. On the contrary, the composition of the jury (four internal instructors and two externals, representatives of SCIC and the interpreting service of the European Parliament), as well as the assessment procedures, which consisted of a recall test and an interview, were totally in line with the recommendations made by major educational bodies (e.g. EMCI, 2011; AIIC, n.d.) and international bibliography on the issue (e.g. Bowen & Bowen, 1989; Gerver,

Longley, Long, Lambert, 1989; Alexieva, 1993; Sunnari, 2002; Monfort et al., 2008). The fact that there was no written assessment included, prior to and in addition to the oral test, owes to functional circumstances. Most candidates came from abroad and, actually, the pool of candidates was rather limited in absolute numbers. Thus, a written assessment might have excluded even more candidates, which would be at the expense of the possibility to sustain a programme with the minimum number of students required (4).

2.2 Identification of problematic areas

What the programme failed to assess sufficiently was a) some of the subjective 'psychological' aspects of training, such as self-control and the ability for team-work, and b) general background knowledge, intellectual curiosity and awareness of current issues. This becomes evident not only through the aptitude test, where the interview mainly revolved around previous education, motivation and future plans of the prospective students (thus leaving behind (b)), but also through the other assessment procedures, i.e. the mid-year and final examinations (including the re-sit) – especially as concerns team-work and booth manners. These aspects have been covered in international bibliography (e.g. Longley, 1989; Lambert, 1992; Moser-Mercer, 1994; Gile, 1995; Mackintosh, 1995; Schjoldager, 1996; Shlesinger, 2000; Donovan, 2003; Sawyer, 2004; Kalina, 2005); albeit, the widespread understanding is that they are the most difficult ones to assess, let alone to predict. This brings the discussion to the criteria which have been used for assessing students in the various assessment stages of the programme.

2.2.1 Skills

The skills tested at all assessment stages after the aptitude test fall under the heading of what Monfort et al. (2008, p. 21) have termed *competence/skill constructs*, drawing on the principles for language testing (Bachman, 2003) and expertise development theory (Hoffman, 1997; Ericsson, 2000; Moser-Mercer et al., 2000). These constructs include skills, such as consecutive and simultaneous interpreting, and sub-skills thereof, e.g. comprehension, memory, task simultaneity, and delivery, and are more or less the same skills and sub-skills on which the EMCI focuses. Nevertheless, as regards the concrete criteria for assessing these skills and sub-skills, there has been quite some confusion. In a nutshell, the criteria stipulated by the EMCI are *accuracy of content* (coherence and fidelity) and *accuracy of form* (grammar, style, and register), *fluency* and *effective-*

ness. These variables are rather ill-defined and often subjective, as Hartley et al. (2003) also assert. A more elaborate set of criteria, which comes closer to what was really applied in the MCI, is provided again by Hartley et al. (2003) in their grid designed for self- and peer-evaluation. From this set of variables, those which were also used consistently in the MCI include: *accuracy* (figures, names, etc.), *cohesion* and *coherence*, *completeness of content*, *terminology* and *register*, *décalage* (normal time distance to original message), *persistence/recovery* (not quitting, error management), and *delivery* (in terms of pace, fluency and communication skills). However, it was never clear how these variables were ranked – if at all – or why variables such as *rhetorical force* (e.g. conveying the intention), *note-taking quality*, *voice parameters* and *booth manners* (mainly team-work) either received sporadic attention, or were not deployed at all.

2.2.2 Rating scale

An issue very closely related to the above problems was posed by the rating scale used for the assessment. The University of Cyprus had to follow the same rating scale for all courses offered, be they under- or postgraduate ones. The grading system is a 0–10 scale, 10 being the highest and 5 being the minimum for successful completion, with 0.5 grades allowed (e.g. 7.5, 8.5 etc.). In some cases, depending on the nature of the course offered, a course can be graded simply on a pass/fail basis. Since it is very hard, if not impossible, to quantify all variables mentioned in the previous paragraph, it has always been difficult to decide, for instance, what level of performance should get a 6 or 6.5 (both within 'good'), or what exactly should make the difference between 8.5 and 9 (both within 'excellent') – a problem common to many other assessment occasions such as the grading of essays. It is perhaps easy to count instances of inaccurate renditions of facts and figures, or count unfinished utterances and instances of severe hesitation, but the fact remains that some other quality aspects such as coherence, critical time distance to original message and communication skills are not quantifiable in the sense of fitting into a strict rating scale. Furthermore, it is extremely difficult to grade persistence, for instance. And even if voice parameters were consistently taken into consideration, the perception of such parameters cannot but be totally subjective (see also the experiments regarding user perceptions of *intonation*, conducted by Shlesinger, 1994 and Collados Aís, 1998).

A similar problem is described by Pöchhacker (2001), who provides a critical discussion regarding the measurement of interpreting performance in the framework of experimental studies in the field. Although this discussion mainly focuses on the lack of comprehensive sets of variables to measure quality, as well

as on the absence of consensus in the interpreting research community regarding a "reliable metric to measure interpreting performance" (Gile in Niska, 1999, p. 120), it is very illuminating both as regards the problem of which variables to use in assessing quality and the problem of quantifying these variables. Thus, the problems posed by the rating scale used in the MCI point to the necessity of revising the practice followed towards a perhaps more concise rating scale with qualitative attributes, instead of quantitative ones, and contrary to the widespread assumptions in favour of refined scoring systems aiming at precise calculations.

2.2.3 Training duration

Finally, another interesting finding which can be read from Table 1 is the rather low success rate in final examinations (averaging 34.4%), as opposed to the very high success rate in re-sit examinations (averaging 91.7%). This gap seems to imply that failure in the final examinations was not due to incapability, but – rather – a matter of *insufficient training time*. Indeed, the final examinations were always scheduled in June, while re-sits were scheduled at the beginning of September, leaving a two-month time over the summer period to prepare. A two-month time is short enough to indicate that a) students merely need some more time to reach a pass, or even higher, level of skill competence, and b) that they are not really irreversibly inadequate in June.

These are the problems which could be identified through a statistical analysis of success rates, as well as through a critical evaluation of the assessment practices followed by the MCI. In the following section, an explanation of the problematic areas will be attempted. It is hoped that this explanation will open up the way for suggesting possible remedies in view of a continuation of the programme in the future.

3. Possible explanations for the problems identified

It is hard to attempt a profound retrospective analysis of possible causes for problems relating to subjective variables, especially since the data gathered reveal a seemingly positive picture of the training offered, as well as a relative adequacy of the assessment procedures, insofar as the inter-subjectively accepted 'formal' assessment criteria are concerned. Such difficulty notwithstanding, it would be worth attempting an explanation of the problems identified, at least for one very important reason; most of these problems are expected to have an impact later, when successful students enter the profession — after all, the most salient reason

for organizing the MCI as a vocational programme was precisely to prepare qualified interpreters for the profession. Thus, a further analysis of problematic areas will benefit the programme as such, but, and perhaps more importantly, it will also be to the long-term benefit of future students. What also needs to be said is that such explanation will not so much provide a thorough exegesis; rather, it will allow going a step deeper into the nature of the problems encountered.

The lack of systematic self-evaluation and peer-assessment procedures with specific criteria and the overall depreciation of team-work and booth manners both on the part of students and of trainers/assessors may initially have been due to personality hindrances of students, though, at a deeper level, it seems to be an issue weighing on the programme itself, as such circumstances should have been anticipated. Apparently, no specific culture of collaboration and appreciation of self-generated progress was cultivated. As a result, no respective consideration was taken during the assessment procedures. However, contrary to the widespread image of interpreting as an 'isolated' and 'individual' activity, the profession relies heavily upon humbleness and good cooperation between colleagues. This should not only be understood in terms of making the professional life easier; moreover, it seems to have a large impact on more technical aspects such as language enhancement and application of techniques, in other words on skill development. When students refuse to learn from peers and remain centred on the teacher's authority (and authenticity) — which is anyway the predominant learning habit cultivated in many domains of contemporary educational systems —, they deprive themselves of the possibility to acquire new language structures and ways of expression or, at least, to learn new tricks from peers who could be more creative in this respect. Apart from stereotypical ways to go around complex syntactic and terminological hitches, every individual seems to be applying more or less conscious idiolectal patterns, from which peers can only benefit.

Thus, the said culture which should be cultivated could have a dual aim: a) to help students overcome a misconceived antagonistic spirit and the initial frustration arising when feedback is given, by highlighting all participants' positive contribution, and b) to lead students to accept the value of all facets of teamwork almost as a necessity, by including their collaborative ability to the assessment criteria used in examinations. As for self-evaluation, although students have actually devoted quite a lot of time practising on their own, it seems that they have not been fully aware of how exactly to assess themselves, or of which concrete assessment criteria to use. In any case, it seems that in terms of feedback, students were not sufficiently exposed to feeling uncomfortable and unfamiliar, as is so often the case in the profession.

This leads to another quite important aspect of the assessment, which concerns the speeches used to assess students' performance. While the level of difficulty

and the topics have consistently been adapted to the respective level of training by all native speakers involved in delivering the speeches, the speakers remained the same during the whole training and all assessment stages. This is due to the fact that it is extremely difficult to have a large pool of native speakers for each language combination, successfully trained to deliver speeches tailored to each of the programme's stages. As a result, students were gradually creating a feeling of familiarity to the pace, style and discourse complexity of the speakers. Moreover, they created a sort of bonding with the speakers, similar to that created with the trainers, which, albeit in isolated cases, allowed students to negotiate the speeches' properties with the speakers. In examinations, this familiarity was obviously enhancing anticipation; at the same time, it was damping the stress created by the exposure to an unfamiliar situation. This prompts us to suspect that the relatively high success rates (cf. Table 2) might not totally correspond to an actual ability of students to cope with real conditions.

The sporadic assessment of background knowledge and knowledge of current affairs was mainly due to two factors: a) background knowledge acquisition is considered to be a continuous process, and b) there was a specially designed course on public speaking and current affairs, precisely in order to account for such knowledge acquisition, among other things. However, the acquisition of background knowledge is not only a cumulative process, but also a skill, which should have been tested. Apart from acquiring information, which inevitably happens when one receives training in conference interpreting, it is perhaps more useful to learn and be able to prove *how* to acquire information (*learning to learn*), i.e. the methodology to do so. This, of course, was part of the daily training routine. But inadequacies and misconceptions never surfaced, since there was no systematic training and these skills were never tested separately.

As to the lack of ranking the rest of the assessment variables, or as to the variables not used in assessment, there was sometimes a divergence of views among different assessors. As was said in the introduction, the assessment panels consisted of the programme coordinators, who were coming from an academic-educational context, the trainers, who were active interpreters, and two representatives from European institutions interpreting services, who were also active interpreters and, usually, were also involved in recruitment tests in their institutions. Since the programme was a vocational one, the judgement given by the active interpreters carried more weight and was decisive. Nevertheless, the active interpreters would put the focus on the overall quality of the final output, without the need to rank the assessment of individual qualities/sub-skills or to explicitly assess sub-skills which are subsumed under other skills (e.g. note-taking in consecutive interpreting). Coming directly from one of the most demanding market contexts (European Union institutions), where the maximum is required of the

output, they would sometimes tend to overlook the difference between skills which require longer experience (e.g. *pace*, *terminology*, *voice control*) and those which need to have been fully trained even after only a crash course (such as *accuracy*, *task simultaneity*, *recovery*, *research methodology*). On the other hand, the assessors coming from an academic-educational context, albeit more familiar with refined assessment practices, were less sensitive to market demands and sometimes tended to appraise the effort more than the result.

The problems concerning the grading system and the insufficient training time are self-evident and do not require any further explanation. However, they will again be discussed, together with all other problems described above, in the framework of the possible remedies/solutions to these problems, which will be suggested in the next section of this chapter.

4. Possible remedies and suggestions for improvement

In order for a) the assessment criteria to be deployed in a more effective way, both as regards their nature and their ranking/refinement, and b) the assessment procedures to respond more effectively to educational needs, it would be useful to cultivate a specific sort of *assessment literacy*. This is a term borrowed from Stiggins (1999 & 2001), and, according to Tsagari (2011, p. 169) is used to describe "the standards of professional excellence that teachers need to attain in relation to assessment such as the ability to critically evaluate, compile, design and monitor assessment procedures […]". Although coming from the context of language learning assessment, this concept may also be of importance to the assessment of interpreting performance in educational contexts. Tsagari (2011, pp. 170–172) offers an extensive overview of problems arising from the incongruence between teachers' practices and recommended best practice and concludes by making a series of suggestions to promote assessment literacy. From these suggestions, the most useful ones in our case seem to be a) the *collaboration of all parties involved in teaching and assessing*, and b) *the involvement of students*.

As to (a), it seems that special meetings between the entities represented in the assessment panels should take place before the start of a programme, in order to reach agreement on the appropriate set of criteria and their ranking by also consulting the advancements in international bibliography. The mere adoption of pre-set, general — and rather vague — educational or professional standards may not necessarily respond to local needs. As to (b), it is imperative to provide students with the agreed set of criteria in order for them to use these criteria during self- and peer-assessment sessions. Students will thus have a solid basis, on which they will produce informed judgements outside classroom contact hours, as well

as an even more targeted preparation for the various examinations. In this way, assessment literacy will be enhanced on the part of assessors and consistently cultivated on the part of students.

Since *learning to learn* seems to be a skill sine qua non for success in the professional life of interpreters, it would perhaps be more meaningful for a training programme to offer a course in *research methodology and public speaking*, instead of in public speaking and current affairs. Knowledge of current affairs is something prospective students should both bring along when applying for a training programme and be willing to exercise on their own. On the contrary, research methodology for interpreters needs to be taught and would aim at familiarizing students with the specific ways in which targeted information on a topic they have no previous knowledge of can be retrieved and managed to serve the needs of an interpreting task, including retrieval of information during the task (cf. Will, 2009) and management of information after the task.

An equally positive effect towards *learning to learn* would be to enhance students' competence with cutting edge technologies in the field of interpreter education (e.g. ICT-technologies and virtual environments — see, for instance, Braun & Taylor, 2011). Working with new technologies will not only prepare students for current advancements in the profession, but, and perhaps more importantly, it will offer them the chance to expose themselves to an invaluable variety of different kinds of speeches and modes of delivery provided by various resources. Such speeches could be used as alternatives to the ones delivered by native speakers, and also as assessment tools, especially for final examinations.

Another important aspect for the assessment of interpreting students would be to introduce the practice of examining two students at a time, in the same booth. Booth manners and efficient cooperation during the task is of paramount importance for the professional life and students need to be exposed to these challenges the soonest possible. Testing students in groups, instead of calling them in separately, will enhance their communicative skills and will expose them to real-life conditions, where *responsibility* is not only to be directed towards the clients, but also towards the colleagues. This is all the more crucial for free-lancers, who cannot always afford the 'luxury' of working with the same colleagues or working within a familiar setting.

Lastly, as to the more practical aspects of training and assessment, any future attempt to continue the MCI should take into consideration that more training time seems to be needed to bring students up to professional standards (see Section 2). A feasible suggestion would be to design a three-semester intensive course, possibly followed by a semester in a foreign university or of work observation (in collaboration with local interpreters). One-year super-crash courses do not seem to be offering suitable preparation for the continuously expanding

demands of the profession both in terms of interpreting mode, as well as in terms of terminological fields. Also, a grading system less dependent on a rigid quantifiability of variables could probably offer a more convenient way of assessing, since, after all, it is always the overall performance which offers the basis for a judgement. The exact nature of such an overall-quality oriented grading system needs extensive analysis and goes beyond the scope of the present chapter. As a preliminary thought, however, a rather simple scale consisting of *good*, *pass* and *fail* could be considered in this respect.

5. Conclusion

This chapter has reported on the educational experience gathered over the three years of implementation of the MCI in Cyprus. Specifically, some statistical data concerning the performance of students over these years have been scrutinized in relation to what they may reveal regarding the assessment criteria and assessment procedures deployed by the MCI. While most of the criteria and procedures were in line with those proposed in relevant bibliography and suggested by professional organizations and services, it was possible to detect some room for improvement concerning both practical issues and issues pertaining to core aspects of assessment tools and criteria. The identification of problematic areas and aspects of the MCI, and the subsequent attempt to explain these, brought to the fore the need for a specific sort of assessment literacy, which might benefit both trainers and students.

In view of a possible continuation or re-conceptualization of the Masters programme in the future, a) the involvement of more assessment criteria, b) the enhancement of the students' responsibility as well as peer- and self-assessment ability in addition to feedback from tutors, and c) the inclusion of advanced technological tools in teaching and assessment are among the aspects which, it is hoped, will prove relevant in meeting the rising demands of contemporary professional life.

References

(AIIC) Advice to students: becoming a conference interpreter. (n.d.). Retrieved December 10, 2011, from http://www.aiic.net/en/tips/students/students7.htm

Alexieva, B. (1993). Aptitude tests and intertextuality in simultaneous interpreting. *The Interpreters' Newsletter, 5*, 8–12.

Bachman, L. (2003). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bartłomiejczyk, M. (2007). https://www.stjerome.co.uk/tsa/abstract/426/ Interpreting quality as perceived by trainee interpreters: Self-evaluation. *The Interpreter and Translator Trainer (ITT)*, *1*(2), 247–267.

Bowen, D. & Bowen, M. (1989). Aptitude for interpreting. In L. Gran & J. Dodds (Eds.), *The theoretical and practical aspects of teaching interpretation* (pp. 109–125). Udine: Campanotto.

Braun, S. & Taylor, J. (Eds.) (2011). *Videoconference and remote interpreting in criminal proceedings*. Guildford: University of Surrey.

Clifford, A. (2005). Putting the exam to the test: Psychometric validation and interpreter certification. *Interpreting, 7*(1), 97–131.

Collados Aís, Á. (1998). *La evaluación de la calidad en interpretación simultánea: La importancia de la comunicación no verbal*. Granada: Editorial Comares.

Directorate-General for Interpretation. (n.d.). Retrieved December 13, 2011, from http://ec.europa.eu/dgs/scic/index_en.htm

Directorate-General for Interpretation and Conferences. (n.d.). Retrieved December 13, 2011, from http://www.europarl.europa.eu/parliament/expert/staticDisplay.do?id=54&pageRank=10&language=EN.

Dodds, M. (1990). On the aptitude of aptitude testing. *The Interpreters' Newsletter, 3*, 17–22.

Donovan, C. (2003). Entrance exam testing for conference interpretation courses: How important is it? *Forum, 1*(2), 17–45.

(EMCI) A university programme at advanced level (Masters-type) in conference interpreting. (2011). Retrieved December 10, 2011, from http://www.emcinterpreting.org/repository/pdf/Core_Curriculum.pdf.

Ericsson, K. A. (2000). Expertise in interpreting: an expert-performance perspective. *Interpreting, 5*(2), 187–220.

Gerver, D., Longley, P., Long, J., & Lambert, S. (1989). Selection tests for trainee conference interpreters. *Meta, 34*(4), 724–735.

Gile, D. (1995). *Basic concepts and models for interpreter and translator training*. Amsterdam & Philadelphia: John Benjamins.

Hartley, T., Mason, I., Peng G., & Perez, I. (2003). Peer and self-assessment in conference interpreter training. Retrieved October 20, 2011, from http://www.lang.ltsn.ac.uk/prf.aspx#lang1.

Hoffman, R. (1997). The cognitive psychology of expertise and the domain of interpreting. *Interpreting, 2*(1/2), 189–230.

Kalina, S. (2005). Quality in the interpreting process: What can be measured and how? In R. Godijns & H. Michaël (Eds.), *Directionality in interpreting – The 'retour' or the native?* (pp. 27–46). Gent: Communication & Cognition.

Lambert, S. (1992). Aptitude testing for simultaneous interpretation. *The Interpreters' Newsletter, 4*, 25–32.

Longley, P. (1989). The use of aptitude testing in the selection of students for conference interpretation training. In L. Gran & J. Dodds (Eds.), *The theoretical and practical aspects of teaching interpretation* (pp. 105–108). Udine: Campanotto.

Mackintosh, J. (1995). A review of conference interpretation: Practice and training. *Target, 7*(1), 199–133.

Monfort, A.-V., Moraki, F., Pouttu, P., & Wang, Q. (2008). Effective screening of candidates for conference interpreting training. Unpublished seminar paper, University of Geneva.

Moser-Mercer, B., Frauenfelder, U., Casado, B., & Künzli, A. (2000). Searching to define expertise in interpreting. In K. Hyltenstam & B. Englund-Dimitrova (Eds.), *Language processing and simultaneous interpreting* (pp. 107–132). Amsterdam: John Benjamins.

Moser-Mercer, B. (1994). Aptitude testing for conference interpreting: Why, when and how? In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the gap: Empirical research in simultaneous interpretation* (pp. 57–68). Amsterdam: John Benjamins.

Niska, H. (1999). Quality issues in remote interpreting. In A. Álvarez Lugrís & A. Fernández Ocampo (Eds.), *Anovar/Anosar estudios de traducción e interpretación* (pp. 109–121). Vigo: Universidade de Vigo.

Pöchhacker, F. (2001). Quality assessment in conference and community interpreting. *Meta, 46*(2), 410–425.

Sawyer, D. B. (2004). *Fundamental aspects of interpreter education: Curriculum and assessment*. Amsterdam & Philadelphia: John Benjamins.

Schjoldager, A. (1996). Assessment of simultaneous interpreting. In C. Dollerup & V. Appel (Eds.), *Teaching translation and interpreting 3: New Horizons*. Amsterdam & Philadelphia: John Benjamins.

Shlesinger, M. (1994). Intonation in the production and perception of simultaneous interpretation. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the gap: Empirical research in simultaneous interpretation* (pp. 225–236). Amsterdam: John Benjamins.

Shlesinger, M. (2000). Evaluation issues in interpreting: A bibliography. *The Translator, 6*(2). Special issue. Evaluation and translation. 363–366.

Stiggins, R. J. (2001). *Student-involved classroom assessment*. Upper Saddle River, NJ: Prentice-Hall.

Stiggins, R. J. (1999). Teams. *Journal of Staff Development, 20*(3), 17–21.
Sunnari, M. (2002). Aptitude tests and selection criteria for interpreting students. *EMCI Workshop proceedings* (pp. 23–26). Retrieved December 10, 2011, from http://www.emcinterpreting.org/resources/simIntoB.php
Tsagari, D. (2011). Investigating the 'assessment literacy' of EFL state school teachers in Greece. In D. Tsagari & I. Csépes (Eds.), *Classroom-based language assessment*. Language Testing and Evaluation 25 (pp. 169–190). Frankfurt am Main: Peter Lang.
Will, M. (2009). *Dolmetschorientierte Terminologiearbeit. Modell und Methode*. Tübingen: Gunter Narr Verlag.

Design and Analysis of Taiwan's Interpretation Certification Examination

Minhua Liu[1]
Monterey Institute of International Studies

Taiwan held its first certification examinations for translators and interpreters in December 2007. Prior to the launch of the *Chinese and English Translation and Interpretation Competency Examinations* (*ECTICE*), a rating scheme was developed and tested in a three-year research project led by this author. The rating scheme, roughly based on the rating mechanism in Carroll (1966), involves the use of two 6-point scales, one for *accuracy* and one for *delivery* – the two criteria used in judging interpretation quality. This chapter discusses the development of the rating scheme and the considerations that went into its design to depart from the multiple-criteria holistic scoring method commonly used in interpretation evaluation. Some issues to be discussed are the comparison of inter-rater reliabilities with other rating methods, such as a proposition-based rating method for accuracy. In addition, a correlation analysis of the two rating criteria – *accuracy* and *delivery* is performed to examine if they can be treated as independent criteria. Raters participating in the *ECTICE* interpretation exams, mostly interpreter trainers who have experience in interpretation evaluation, are surveyed about their views on this rating scheme and their experience in using the rating scales.

Key words: accuracy, certification, delivery, interpretation, rating.

1. Introduction

Taiwan held its first certification examinations for translators and interpreters, *Chinese and English Translation and Interpretation Competency Examinations* (hereafter, *ECTICE* exams), in December 2007. Prior to the launch of the exams, Taiwan's Ministry of Education, the certifying body, commissioned a three-year research project, in which this author and her team proposed a plan for the structure of the certification program, developed test specifications and rating schemes for the written translation tests and consecutive interpretation tests. This chapter describes the development of the rating scheme for the consecutive interpretation exam (hereafter, *ECTICE* interpretation exam) and the considerations that went into its design, a departure from the multiple-criteria holistic scoring method commonly used in interpretation evaluation. Some issues to be discussed are the comparison of inter-rater reliabilities with other rating methods, such as a proposition-based rating method for accuracy. In addition, a correlation analysis

1 mliu@miis.edu

of the two rating criteria – *accuracy* and *delivery* – is performed to examine if they can be treated as independent criteria. Raters participating in the *ECTICE* interpretation exam, mostly interpreter trainers who have experience in interpretation evaluation, are surveyed about their views on this rating scheme and their experience in using the rating scales.

Considering that the *ECTICE* exams are high-stakes national exams that attract a large number of test-takers,[2] several considerations went into designing the rating scheme to make it a standard evaluation method that is 1) less subject to arbitrariness from subjective judgment, 2) accurate and reliable, and 3) simple and feasible.

2. ECTICE interpretation exam

The *ECTICE* interpretation exam is composed of an English competency paper-and-pencil test and two interpretation tests, with the former serving as a screening mechanism for the second-phased interpretation tests. The interpretation tests include a short consecutive interpretation (short CI) test and a long consecutive interpretation (long CI) test.[3] Both performance tests are for generalists. Each test further consists of two English passages and two Chinese passages to be interpreted into Mandarin and English respectively. The length of each passage in the short CI test is about 3 minutes and the length of each passage in the long CI test is about 5 minutes. Each short CI passage is divided into six to eight interpretation segments and that of the long CI into two to three segments (Ministry of Education, 2009b). The topics covered in both short and long CI tests are of a non-technical nature. A brief summary of the content and some difficult terms and their equivalents are provided for the test-takers.[4] The tests are administered to a small group of test-takers in a language-lab setting. Test items are presented aurally and test-takers' interpretations are recorded digitally and turned into individual audio files. The time allowed to interpret each interpretation segment is limited and pre-set.[5]

2 The numbers of test-takers of the *ECTICE* translation exam are 706 in 2007, 303 in 2008, 501 in 2009, 450 in 2010, and 393 in 2011, and those of the *ECTICE* interpretation exam are 395 in 2007, 143 in 2008, 244 in 2009, 230 in 2010, and 209 in 2011 (Ministry of Education, 2011).

3 Taiwan's simultaneous interpretation market is small and is dominated by a small group of professionals. The necessity of adding a simultaneous interpretation test is still under review by the Ministry of Education.

4 Terms provided include formal names (e.g., institutions, organizations, titles, etc.) and jargon, which can be easily looked up on the Internet, to which test-takers are not allowed access.

5 The pre-determined time for interpretation for each test and each interpreting direction was tested in several pilot studies to ensure test-takers have a reasonable amount of time to complete an interpretation segment.

The test-writers of the *ECTICE* interpretation exam are all professional interpreters and interpreter trainers. They are provided with a small booklet of test-writing guidelines including suggestions on subject areas, level of difficulty, considerations for ways to judge and control test difficulty, suggestions for revising base texts, and instructions for writing a summary of a text. Two senior interpreter trainers choose among four sets of tests and decide on one set for each of the short and long CI tests. The chosen set of tests are then recorded as audio files, with the speech rate set at 100 to 110 English words per minute or 160 to 175 Chinese characters per minute (Ministry of Education, 2009b).

2.1 Rating criteria and development of rating scales

During the process of selecting the rating criteria for the *ECTICE* interpretation exam, we first operationalized the construct of interpretation competency as the ability "to adequately comprehend the message conveyed in a typical talk at an occasion where consecutive interpretation service is needed and can accurately convey the message into another language using appropriate grammar and word choices and deliver the message in a smooth and easy-to-understand manner" (Liu, Chang, Lin, Chen, Yeh, & Luo, 2005, p. 45; Ministry of Education, 2009a, p. 2). We decided on selecting the most important criteria documented in the interpretation literature. Content fidelity (including accuracy and completeness, the two key components frequently referred to under fidelity) and output quality (including intelligibility of the message, smoothness of delivery, and appropriateness of language use in the target language) are the two criteria that are most widely used in interpretation rating practice (see Liu, Chang, & Wu, 2008 for a discussion of the rating practice of 11 interpreter training programs). For the purpose of labeling, these two criteria are called *accuracy* and *delivery*.

Carroll (1966) discussed a rating method developed for the purpose of judging the quality of machine-translated pieces. There are two rating criteria used in this method, *intelligibility* and *informativeness* (representing fidelity or accuracy). *Intelligibility* of a translated sentence is described as reading "like normal, well-edited prose and be readily understandable in the same way that such a sentence would be understandable if originally composed in the translation language" (p. 57). The *intelligibility* scale has 9 levels (1 to 9), with 9 being the highest score. An *intelligibility* scale descriptor can read like this, "The general idea is almost immediately intelligible, but full comprehension is distinctly interfered with by poor style, poor word choice, alternative expressions, untranslated words, and incorrect grammatical arrangements. Postediting could leave this in nearly acceptable form" (Level 6) (p. 58). *Informativeness* is used to judge if translated sentences "twist,

distort, or controvert the meaning intended by the original" as little as possible (p. 57). The *informativeness* scale has 10 levels (0 to 9) and 9 is the lowest score. The reason behind this reverse scaling is that, in judging *informativeness*, the original sentences are judged, relative to the translated sentences, if they "contain no information that would supplement or controvert information already conveyed by the translation" (p. 57). An *informativeness* scale descriptor can read like this, "By correcting one or two possibly critical meanings, chiefly on the word level, it gives a slightly different 'twist' to the meaning conveyed by the translation. It adds no new information about sentence structure, however" (Level 3) (p. 58).

During the process of developing the scales for the *ECTICE* interpretation exam, we chose to adopt the two rating criteria in Carroll's rating method but changed the names of the criteria to *delivery* and *accuracy*. The term *delivery* was chosen to reflect the different dimensions a rater needs to consider when judging an interpretation performance. The descriptors in Carroll's *intelligibility* scale are multidimensional as they contain elements related to both ease of understanding and language use (including style, word choice and grammar) (p. 58) (see, for example, the descriptor of Level 6 above). Carroll explained that since the descriptors reflect the actual quality of the 200 translated sentences used in developing the descriptors, the raters should be able to make reliable judgments on this criterion (p. 58). For the purpose of evaluating interpretation performance, the same consideration for a multidimensional scale is also necessary because an interpretation output is usually judged on how easily the message can be understood, how smooth the delivery is, and if target language use is appropriate. As to the choice of *accuracy*, since we decided to reverse the scaling of Carroll's *informativeness* scale to avoid confusion during scoring, we found the label *accuracy* to be clearer in reflecting how an interpretation output compares to the original in term of content fidelity. Another consideration is that the final scores of the *ECTICE* interpretation exam are calculated by adding the *accuracy* and *delivery* scores (see 2.4 for more details). The total score from two reversed scalings will not be able to provide information on a test-taker's performance.

In several small pilot rating sessions, we tested Carroll's original 9-point *intelligibility* scale and the 10-point *informativeness* scale and compared the results with those from 7-point scales and 5-point scales. Based on the rating results and the feedback we received from raters at these rating sessions, we decided on using the 5-point scale based on the following reasons: 1) Human translations generally have better quality than machine translation and thus may not need as many levels in the scale as Carroll's;[6] 2) Carroll's *informativeness* scale contains a

6 Carroll reported that machine-translated sentences are much more variable in *intelligibility* and *informativeness* than human-translated ones (p. 62).

couple of levels that are merely described as "between level Y and level Z" without an actual descriptor; 3) our raters expressed that it was difficult sometimes to distinguish between two seemingly very similar levels when judging the sample interpretation performance and this was true for both scales (Chen & Liu, 2007)[7].

We continued to test the two 5-point rating scales in several larger pilot rating sessions, after which the wording of some descriptors were modified a couple of times to reflect the actual performance in the samples tested in the pilot studies. When the two scales were used in the first *ECTICE* interpretation exam in 2007, more adjustments to the scales were deemed necessary, as it was shown that the range of the interpretation competency of the test-takers was wider than the participants in the pilots, who were mostly graduate-level interpretation students or upper-level undergraduate students who had taken interpretation classes. An extra level '0' was then added to the *accuracy* scale to represent the situations where no interpretation was rendered for a particular source language section (see Appendix 1 for the two scales used in the *ECTICE* interpretation exam).

2.2 Rating units

Carroll's original design of the rating method uses individual sentences as rating units to allow "a substantial number of relatively independent judgments" to be obtained on a translation (p. 56). He considers that individual sentences can "convey at least the 'core' meaning" in the original sentences (p. 56). However, considering the nature of consecutive interpretation and the difficulty to match individual target language sentences with source language sentences, we decided to use a segment of several sentences that cohesively forms an idea as the rating unit. The actual length of the segments depends on how many sentences it takes to form a self-inclusive idea. In a typical short CI passage, there are usually 6 to 8 rating units, corresponding to the interpretation units in each passage. In a long CI passage, there are also usually 6 to 8 rating units, divided among the two to three interpretation segments (Liu, Chang, Chen, Lin, & Wu, 2007).

7 Among studies that also used Carroll's scales (e.g., Clifford, 2005; Tiselius, 2009), Tiselius also reduced the number of levels to 6 in both *intelligibility* and *informativeness* scales for more efficient rating of spoken language (p. 101, 104).

2.3 Rater training

Raters involved with the *ECTICE* interpretation exam are all interpreters, most of whom are interpreter trainers who have experience in evaluating interpretation performance in class or in exams. English into Chinese tests are all rated by raters with Mandarin as their native language, and Chinese to English tests are by raters with English as their native language or by those who have equal command in both English and Mandarin. All raters go through a four to five-hour training session before starting to rate.

Because the *accuracy* scores depend heavily on the number of important meaning units missed or incorrectly interpreted, a small group of three to four interpreter trainers discuss and reach a consensus on the important meaning units in each rating unit, which are then marked on the source texts.

After the last testing session of the exam ends, the same group of interpreter trainers review samples from the test-takers' interpretation recordings and select examples that they think exemplify the different levels on the *accuracy* scale or the *delivery* scale. At least five to six examples representing different styles of interpretation output in each language direction are selected for each level in each scale. The examples are chosen with the aim of providing a heterogeneous mix of interpretation performances.

The rater training session for each language direction is conducted separately after a joint session where a briefing is given on the specifics of that year's *ECTICE* interpretation exam, the rating scales, and the rating procedure. Two of the aforesaid interpreter trainers serve as rating trainers, one for each interpreting direction. After the joint session, the raters in each group first listen to the recordings of the source speeches in that interpreting direction. In order for them to develop an understanding of the difficulty level of each speech on factors such as speed, information density, terms and background knowledge, the raters are encouraged to take notes while listening to the speeches as if they are to interpret. Printed scripts of the source speeches are provided to all raters who then discuss the important meaning units marked on the texts (see Appendix 2 for a sample rating sheet with important meaning units marked). After all raters reach an agreement on the important meaning units, examples previously selected for each level of the scales are presented to the raters to help them develop a stronger understanding of the scales. The rest of the rater training session is spent on rating practice using samples from that year's test-takers' interpretation recordings and on discussion of rating results.

The arrangement of raters is made in such a way that every test-taker's English to Chinese interpretation is rated on *accuracy* by two raters and on *delivery* by another two raters. The same arrangement is made for the Chinese to English interpretation. This way, each test-taker's interpretation performance is judged by a total of eight raters.

Raters work independently without knowing the scores given by other raters. Raters who judge *accuracy* listen to the recordings of the test-takers and check accuracy against the scripts of the source speeches, on which the agreed-upon important meaning units in each rating unit are marked. Each rating unit (consisting of several sentences forming a coherent message) is given a score from 0 to 5 on the *accuracy* scale. Raters who judge *delivery* listen to the recordings without comparing them to the transcripts of the source speeches. Each rating unit is also given a score from 0 to 5 on the *delivery* scale. To determine the score of a passage, the scores of all rating units (usually six to eight) in each rating criteria (*accuracy* or *delivery*) are added up and converted into percentage scores. If there is a discrepancy of 10 points or more between two raters' scores, a third rater is asked to rate the test. The final score for each language direction in each test (short CI or long CI) is calculated by adding and then averaging the *accuracy* scores and the *delivery* scores. The final score of each test (short CI or long CI) is determined by adding and averaging the scores of the two interpreting directions. Test-takers have to obtain a passing score of 80 for each test (short CI or long CI) to be granted a certificate. But a passing grade in one test remains valid for up to three years if test-takers decide to pursue this certification again.

3. Raters' response to the rating scheme

In an effort to establish face validity of the *ECTICE* interpretation exam and to understand raters' experience in rating this exam, we conducted surveys on raters for two consecutive years (Liu et al., 2007; Liu, Chang, Chen, Lin, Lee, & Chiu, 2008). The feedback revealed that the tests are generally considered to be a good measure of interpretation ability. As a group, they found separating the two rating criteria and having them independently judged by two groups of raters meaningful, as they often encounter faithful interpretations badly delivered or smoothly delivered speeches that are far from the original in content.

As for dividing the passages into smaller rating units, those raters surveyed generally think that it is a good practice because it allows them to examine the interpretation output more objectively by giving six to eight scores instead of one

holistic score, which oftentimes just reflects a general impression of the whole interpretation performance. They did express the need to spend more time on rating because of the closer examination of the interpretation output in each rating unit and of having to decide on six to eight scores.

The raters unanimously agreed that the rater training sessions greatly helped them interpret the scores and link the descriptors on the scales to actual interpretation performance. They also expressed the need for more examples representing a bigger variety of test-taker performances, as they still sometimes had difficulty in judging performances falling between two levels, particularly between levels 3 and 4. They also indicated that such difficulties were encountered more when judging *delivery*. Comparatively, judging *accuracy* seemed to be much more straightforward, thanks to, according to the raters, the markings of important meaning units and the thorough discussions raters had to reach an agreement on the important units. The raters also made it clear that leaving sufficient time for discussions among raters is very important for them, not only in the actual rating but also psychologically, as they felt that their judgment would be more in agreement with that of the other raters. They also suggested leaving more time for rating practice and ample time for discussing the results.

One question that arose in the survey was how closely the two skills of short CI and long CI are related. Raters whom we surveyed mostly agreed with the design of the *ECTICE* interpretation exam[8] and thought that short CI and long CI do not involve exactly the same set of skills and thus warrant different testing procedures. However, the correlation coefficients obtained in our analysis showed that the test-takers' performance in short CI and long CI correlated at .76 ($p <$ .01) in English to Chinese and .52 ($p < .01$) in Chinese to English. These results indicate that the two skills seem to overlap to some extent and that test-takers' performance seem to be quite consistent from English to Chinese. The lower correlation between the two tests from Chinese to English may reflect a lack of consistent performance of the test-takers or may also indicate bigger variation in rating when a test-taker's weaker language is assessed. The question of whether both short CI and long CI tests are needed to assess skills in consecutive interpretation remains to be answered.

One interesting issue emerging from the feedback from the raters is the relation between the two criteria. Many raters expressed that it is sometimes difficult to distinguish if a problem in the interpretation output is an accuracy or a delivery issue. For example, a wrong choice of terms can be judged as an accuracy problem but also a language one. Likewise, bad grammar, a language problem according to the descriptors on the *delivery* scale, can cause miscomprehension

8 On a Likert scale of 1 to 5 indicating the level of agreement, the raters' score averaged 4.83.

of the source speech, which becomes an accuracy problem. This is particularly true when judging the interpretation output in the test-takers' B language, which, in the case of over 99% of the test-takers' of the *ECTICE* interpretation exam, is English. In addition, the raters were concerned that judgment difficulty will result in a reverse halo effect in the test-takers' scores. This situation prompted us to examine the correlation between the *accuracy* and *delivery* scores.

4. Correlation of the two criteria

Carroll (1966) originally considered *intelligibility* and *informativeness* "conceptually separable variables" (ibid, p. 57). However, the results of the experiments showed that the criteria were highly correlated. Similar results are also found in Clifford (2005) where *intelligibility* and *informativeness* are correlated at .746 ($p<.000$), and where *intelligibility* and *informativeness* are correlated with a third criterion, *style*, at .690 and .691 (both at $p<.000$) respectively in a simultaneous interpretation test. Clifford (2005) suggested that it is not appropriate to treat *intelligibility* and *informativeness* as two separate criteria and that performance tests assessed this way cannot be considered mulitidimensional, but unidimenstional, i.e., only measuring one construct.

When we piloted our rating method, we also tested if the two criteria are correlated. The correlation coefficients of the *accuracy* and *delivery* scores are .668 ($p=.000$) in the English to Chinese group and .743 ($p=.000$) in the Chinese to English group (Yeh & Liu, 2006), both considered to show moderate correlation and substantial relationship (Guilford, 1973). Further research should show if there is a halo effect or a reverse halo effect and if the two criteria should remain to be evaluated separately.

5. Using monolingual raters to judge accuracy

Carroll (1966) used two groups of raters in his experiment to study the effect of source language knowledge on rating. One group was composed of English speakers with expertise in reading scientific Russian (called the 'Russian readers'). The other group was composed of high-performing science majors with high verbal intelligence and no knowledge of Russian (called the 'monolinguals'). The Russian readers judged the translated sentences against the original source sentences, while the monolinguals compared the translated sentences to a translation done by experts. The results showed that monolinguals achieved signifi-

cantly greater reliability in both their *intelligibility* and *informativeness* ratings than the group of Russian-reading raters.

This particular result of Carroll's study is not only theoretically interesting but also has practical implications. It is particularly relevant when the pool of raters is small. This is the case with the *ECTICE* interpretation exam as the population of our 'ideal raters' (i.e., professional interpreters and interpreter trainers) is not very big. We replicated this part of Carroll's study in a small study, where we investigated the possibility of having the raters judge interpretation accuracy without reading the original script. We had two groups of raters, all interpreters, judging the same interpretation samples on accuracy. One group used the script of the original speech and the other group used an interpretation version provided by a professional interpreter. Both groups of raters used the same *accuracy* scale mentioned above. We found similar results as those of Carroll's study. The inter-rater reliability was higher in the group using the model interpretation (.812, p=.000) than the group using the original speech (.767, p=.000) and the scores of the two groups were highly correlated at .916 (p=.000), indicating that the two methods are highly interchangeable (Yeh & Liu, 2006)[9].

Both Carroll's and our results imply that knowledge of the source language may not be a prerequisite for assessing the accuracy of a translation (or interpretation) and that it is highly feasible to use a good version of translation (or interpretation) as an alternative of the original text (or speech) when raters with good verbal skills are involved. This is promising, as in the case of a large-scale examination that attracts a great number of test-takers, it may be difficult to find professional interpreters or people with high skills in both languages as raters. In this case, having monolingual raters use a model interpretation may be a good alternative. However, we also need to consider the practicality of this practice in real interpretation exams. Creating a model interpretation may be more difficult than producing a model translation and the quality of the model interpretation can affect the results of the rating. In addition, this method may meet more skepticism and resistance from the interpretation profession.

9 Different results are shown in Tiselius (2009), where interpreter raters using the original to grade had a higher inter-rater reliability (.65) than non-interpreter raters who used a translation of the original (.50). Tiselius attributed the higher correlation among interpreter raters to their similar background (p. 115).

6. Judging fidelity: Scale-based rating vs. proposition-based rating

In one of our pilot studies to test reliability of our *accuracy* scale, we compared the results of using the *accuracy* scale with those of a proposition-based rating,[10] where we calculated the percentage of propositions of each source text correctly interpreted (Liu & Chiu, 2009). We checked how closely each transcribed interpretation matched against the propositions of each source text. A score of 1 was given when the meaning of a proposition was correctly interpreted. Otherwise, a score of 0 was given. Two Chinese native speakers with graduate-level interpretation training served as raters. The two raters first did a trial rating session individually using the interpretation of three randomly chosen participants. After discussing the results of the trial rating, they agreed on some principles on rating and then proceeded with the rating. To assure better consistency in rating, all participants' interpretation of the same section of a particular speech was rated before proceeding to the next section. The final score for each interpretation was the average of the scores given by the two raters, calculated by dividing the number of correctly interpreted propositions by the total number of propositions in each source material. To test how the two ratings correlated, we first converted the *accuracy* scores from the scale-based rating to percentage points. Pearson correlation coefficient showed that the two achieved very high correlation at .945 (p=.000). This is very encouraging in that the rather easy-to-do scale-based rating for fidelity can be used as a substitute for the highly rigorous yet extremely tedious proposition-based rating when judging accuracy in interpretation.

7. Discussion and conclusion

The *ECTICE* exams (including both translation and interpretation exams) have been held five times since 2007. As a government-sponsored national exam, the *ECTICE* exams have attracted much attention from mostly young Taiwanese who aspire to become professional translators and interpreters. However, despite the clearly stated goal of assessing professional competency,[11] the examination practice (e.g., choice of test items, general test difficulty level, etc.) and standards (in

10 A proposition is the smallest unit that carries a meaning. A typical proposition is composed of a predicate and one or more arguments. The predicate specifies the relationship between the arguments (Kintsch & van Dijk, 1978).

11 *Directions Governing the ECTICE Exams*, an official government document, state the goal of the exams as "assessing the Chinese and English translation competency of people who wish to engage in professional transition work" (Ministry of Education, 2009a, p. 1).

terms of the use of the rating scales and what is conveyed to raters at rater training), and the rather low passing rate of around 10%, many test-takers and people in the field of translation and interpretation do not seem to view the *ECTICE* exams as true certification exams for professionals. This is partly due to the fact that all graduate-level and many undergraduate-level translator and interpreter training schools in Taiwan have their own exit exams, where a successful candidate is considered to possess the necessary skills to become a professional translator or interpreter, hence the name 'professional exams' used in many schools. These exit exams usually have a component of testing simultaneous interpretation skills, which are not tested in the *ECTICE* interpretation exam. It is particularly for this reason that some schools do not see the *ECTICE* interpretation exam as a full-fledged examination for professional interpreters. As mentioned earlier, the market for simultaneous interpretation (conference interpretation) is quite small in Taiwan and is dominated by a small group of professional interpreters. In contrast, the market for consecutive interpretation (e.g., in-house interpreters working for public or private institutions, interpreters who serve the vibrant business and industrial sectors, and interpreters who work at legal settings) still has room for growth and it is the interpreters who work in this market that the *ECTICE* interpretation exam is targeting. A limitation with the exit examinations at translator and interpreter training schools is that they are restricted to the students of these schools. A national exam is thus necessary to serve the general public.

During the five years since the *ECTICE* interpretation exam first took place, we have observed the adoption of the rating scales and the rating practice by several translator and interpreter schools for their exit exams (Liu, Chang, & Wu, 2008) and some schools even considered adding the *ECTICE* interpretation exam to the requirements for graduation. It was explained to us that there had been a lack of consistent standards and assessment practice among these training schools and an effort to standardize was welcomed. Indeed, assessment in interpretation has been characterized by arbitrary selection of test content, a lack of consistent test administration practices, and a failure to establish and respect objective scoring criteria (Sawyer, 2000). Our survey of 11 interpreter training schools on their exit exams showed that most schools use an analytical scoring method that evaluates multiple dimensions of interpretation performance. However, the criteria used for judging those dimensions are often only labeled but not described. Individual scores for each criterion are either not reported or not used toward calculating the final score, which oftentimes represents a rater's overall impression of a performance. Rating may be done independently, but the final decision of pass or fail is often based on joint agreement among raters. Professional interpreters serve as raters in most of these schools and it is often their expertise that is relied on when judgment is made because there is usually no rater training (Liu, Chang, & Wu, 2008).

We hope that our attempt to create a standardized way of assessing interpreter competency in the *ECTICE* interpretation exam not only serves the purpose of a certification in setting standards, developing professional practice, and improving user protection, but can also inform training schools in the practice of their exit exams and help promote a more reliable way of assessing interpretation skills at the professional level. For this purpose, we are proposing the following suggestions as good practice:

1. If a holistic scoring method is used, lay out specific criteria with clear descriptors even though scores are not assigned to each criterion.
2. If an analytical scoring method is used, in addition to clear descriptors for all criteria, how the sub-score of each criterion (including its weighting, if any) and the total score are calculated should be made clear.
3. Divide the source text into smaller rating units to allow more precise decisions when judging fidelity.
4. Have clearly differentiated important and secondary meaning units or clearly defined errors when evaluating fidelity of interpretation.
5. Rater training is useful in making raters feel more assured of the whole rating process. It is particularly important to provide samples with sufficient varieties for each level on a rating scale. Giving raters adequate time for rating practice and discussion can also help raters achieve ease and consistency during actual rating. However, how much rater training contributes to achieving reliable and valid ratings is not clear and requires further research,

References

Carroll, J. B. (1966). An experiment in evaluating the quality of translations. *Mechanical Translation,* 9, 55–66.

Chen, B., & Liu, M. (2007). A more objective approach to translation evaluation: Evaluation scales and evaluation units (in Chinese). *Journal of the National Institute for Compilation and Translation*, *35*(3), 55–72.

Clifford, A. (2005). Putting the exam to the test: Psychometric validation and interpreter certification. *Interpreting, 7*(1), 97–131.

Guilford, J. P. (1973). *Fundamental statistics in psychology and education* (5th ed.). New York: McGraw-Hill.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363–394.

Liu, M. Chang, C-C, Chen, T-W., Lin, C-L., & Wu, S-C. (2007). *Establishment of an evaluation mechanism for the consecutive interpretation tests in Taiwan's Translation and Interpretation Competency Examination – 1st phase* (in Chinese). Unpublished technical report. Taipei: National Institute for Compilation and Translation.

Liu, M., Chang, C-C, Chen, T-W., Lin, C-L., Lee, C-C., & Chiu, Y-H. (2008). *Establishment of an evaluation mechanism for the consecutive interpretation tests in Taiwan's Translation and Interpretation Competency Examination – 2nd phase* (in Chinese). Unpublished technical report. Taipei: National Institute for Compilation and Translation.

Liu, M., Chang, C-C., Wu, S-C. (2008). Interpretation evaluation practices: Comparison of eleven schools in Taiwan, China, Britain, and the USA (in Chinese). *Compilation and Translation Review, 1*(1), 1–42.

Liu, M., Chang, W-C., Lin, S-H., Chen, B., Yeh, S-P., & Luo, H. (2005). *Establishment of evaluation standards for translators and interpreters in Taiwan* (in Chinese). Unpublished technical report. Taipei: National Institute for Compilation and Translation.

Liu, M., & Chiu, Y-H. (2009). Assessing source material difficulty for consecutive interpreting: Quantifiable measures and holistic judgment. *Interpreting, 11*(2), 244–266.

Ministry of Education (2009a). *Directions governing the Chinese and English Translation and Interpretation Competency Examinations*. Taipei: Author.

Ministry of Education (2009b). *Guidelines for the Chinese and English Translation and Interpretation Competency Examinations*. Taipei: Author.

Ministry of Education (2011). *Chinese and English Translation and Interpretation Competency Examinations*. Retrieved December 20, 2012 from http://www.edu.tw/bicer/content.aspx?site_content_sn=8493

Sawyer, D. B. (2000). Towards meaningful, appropriate, and useful assessment: How the false dichotomy between theory and practice undermines interpreter education. *ATA Chronicle,* 29(2), 32–40.

Tiselius, E. (2009). Revisiting Carroll's scales. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies* (pp. 95–121). Amsterdam: John Benjamins.

Yeh, S-P., & Liu, M. (2006). A more objective approach to interpretation evaluation: Exploring the use of scoring rubrics (in Chinese). *Journal of the National Institute for Compilation and Translation, 34*(4), 57–78.

Appendix 1

Rating scales for *accuracy* and *delivery* of Taiwan's ECTICE interpretation exam

The Accuracy Scale

Level	Description
5	The message in the interpretation is the same as that in the original speech. It contains no errors.
4	The message in the interpretation is similar to that in the original speech. It contains one or two minor errors.
3	The message in the interpretation is slightly different from that in the original speech. It contains one major error or many minor errors.
2	The message in the interpretation is very different from that in the original speech. It contains two or more major errors.
1	The message in the interpretation is completely different from that in the original speech.
0	No interpretation is produced.

The Delivery Scale

Level	Description
5	The interpretation is fully comprehensible and very coherent with few instances of hesitation, repetition, self-correction, and redundancy. It contains few inappropriate usages of grammar or terms.
4	The interpretation is mostly comprehensible and coherent with some instances of hesitation, repetition, self-correction, and redundancy. It contains some inappropriate usages of grammar or terms.
3	The interpretation is generally comprehensible but is not very coherent and has many instances of hesitation, repetition, self-correction, and redundancy. It contains many inappropriate usages of grammar or terms.
2	The interpretation can be understood with great difficulty.
1	The interpretation cannot be understood at all.
0	No interpretation is produced.

Appendix 2

Sample rating sheet (with important meaning units highlighted) of Taiwan's ECTICE interpretation exam

Unit	Original sentences	Accuracy 0-1-2-3-4-5
01	I hope to do two things today: explain why I believe the media's role in increasing visibility for HIV/AIDS is so important, and what I believe is needed to stop this epidemic.	
02	The media have played a crucial role in highlighting the most important issues of our time. Yet HIV/AIDS may be the greatest challenge of all. As I have become more engaged in global health issues over the past decade, one thing has become clear to me, that is, not enough is being done about the millions of preventable deaths each year from diseases like AIDS.	
03	In part, that's because people aren't aware of what is happening. We don't see these issues covered enough in newspapers, radio and television. People need to see the problems up close, to act. That is why raising the visibility of these issues and passing on accurate health messages are so incredibly important.	
04	You are vital in the fight against AIDS for three reasons: First is attention. You have the power to bring greater attention, action, and co-operation from government leaders. The second is Stigma. You can eradicate the stigma and discrimination associated with AIDS by providing accurate information and humanizing media coverage.	
05	The third is Information. Even the best, most compelling data of scientists and health experts reaches a limited audience. You deliver practical, life-saving information to people in the hardest-hit areas that will help them protect themselves and connect with social services.	
06	So what is needed from the media? To win this battle, I believe three critical things are needed. First is visibility. We need to keep up the pressure through media coverage. Spread the word throughout your newsrooms and media organizations that this is THE story—the most important story of our time.	
07	Second, we can't just tell people about the problems. We have to tell them about effective, affordable solutions—and about how little money it takes to save a life. The third critical element is leadership. Encourage your peers to do more. If every media company in the world took on HIV/AIDS, just think of the progress we could make.	

Certification of Social Interpreters in Flanders, Belgium: Assessment and Politics

Britt Roels[1]
Central Support Cell for Social Interpreting and Translation, Junction Migration-Integration, Belgium

This chapter deals with the certification procedure for social interpreters (SI) in Flanders. It sketches the development phases of an objective, reliable and valid assessment procedure. An SI competency profile issued by the Flemish government serves as the basis for the development and the final assessment. There are three main development phases: the analytical, the experimental and the finalizing phase. The final exam format consists of four evaluative parts: a Dutch and foreign language proficiency test and two role plays. This format is described meticulously. This chapter also highlights a number of quality assurance mechanisms in place when developing assessment material. Moreover, several fair assessment measures are available to assure the equality and equity for all SI candidates. Additionally this chapter sheds light on the political and historical background of the Flemish SI sector and its uncertain future.

Key words: social interpreter, certification, validation, assessment, Flemish politics.

1. Social interpreting in Flanders

The federal state of Belgium comprises three autonomous regions. Flanders covers an area of 13,684 km² and has a population of about six million. Flanders' official language is Dutch. The past three decades the Flemish government has been developing a broad policy for civic integration due to the emerging reality of a multilingual society. An element of this policy is the provision of social interpreters (SI) to immigrants. The concept of social interpreting is only used in Flanders and refers to the above-mentioned federal structure of Belgium (Vermeiren et al., 2009, p. 298).

Social interpreting is defined as the faithful, complete and neutral transfer of oral messages from a source language into a target language in the sphere of public and social care. Social interpreting is community interpreting but excludes interpreting in the legal, police and asylum contexts. It covers both interpreting in face-to-face situations and provided over the telephone (SERV, 2007).

Traditionally, social interpreting was done by non-professionals such as family or friends. It was not considered a profession but merely a service to friends or members of one's own linguistic group. Although these non-professionals pro-

1 britt.roels@kruispuntmi.be; britt.roels@gmail.com

vided a valuable service, there was no guarantee that the interpreting was done in a reliable and correct manner (Vermeiren et al., 2009, pp. 299–300).

A societal awareness of the role of interpreters with the purpose of ensuring equal access to social services, gradually emerged. This resulted in the creation of some pioneer social interpreting agencies in the 1980s, 1990s and 2000s (COC, 2007). Currently, there are 9 social interpreting agencies. Eight are local and one is centralized. They are either non-profit or governmental organizations (Vlaams Parlement, 2009).

In 2004, the Central Support Cell for Social Interpreting and Translation (COC), was founded to support, develop and harmonize this relatively new and heterogeneous sector. Its main tasks are organizing SI courses and certification exams based on a number of quality standards agreed upon by the entire sector (COC, 2007). Additionally, a *2009 Integration Decree* issued by the Flemish government, declares that SI agencies should no longer dispatch uncertified SIs to interpreting assignments if certified SIs exist and are available (Vlaams Parlement, 2009).

2. The Flemish government values competencies

In 2004 the Flemish government issued a *Decree concerning the acquisition of a title of professional competency*. This decree ratifies a system to validate and certify a person's competencies in order to execute a certain profession, regardless of the fact whether he[2] developed these competencies formally or informally (Vlaamse Gemeenschap, 2004).

After consulting the Social Economical Council of Flanders (SERV)[3], the Flemish government determines the professions for which these titles can be distributed. Hereafter, the SERV writes the relevant competency profiles. A *competency profile* (CP) delineates a profession, its working conditions and the necessary competencies (Vlaamse Gemeenschap, 2004). A CP is an instrument that can either be used by professionals to guarantee the quality of their profession, or by vocational and educational institutions as a reference to formulate the final goals of their courses.

As a final step, a *Standard* containing the indispensable competencies necessary to execute the profession, is derived from the CP. The *Standard* also formulates conditions regarding the assessment procedure for test centers (Vlaamse Gemeenschap, 2004). Any institution can apply to become a test center, if it possesses the necessary knowledge relating to the profession. Currently, there are 54 test centers and more than 200 CPs.

2 With all due respect for the readers and for purely practical reasons, I will use the male grammatical gender for a social interpreter throughout this chapter.

3 The Social Economical Council of Flanders (SERV) advises the Flemish Parliament and the Flemish government about all economic and social matters that lie within the authority of Flanders.

In 2007, the SERV published in collaboration with the SI sector, an SI competency profile (SERV, 2007). In 2008, the COC became the official SI test center.

3. The COC as a social interpreter test center

In 2009 and 2010 the European Social Fund (ESF)[4] allotted funding to the COC to organize 300 certification exams. The format of the exam employed in this period was the result of a seesaw process. This process had started with the COC's inception and thus before the existence of the *Standard*. Hence the format of this exam was not exactly conform with the *Standard*. The exam consisted of 5 parts: a Dutch and foreign language (FL) proficiency test, a note-taking reproduction exercise, a sight-translation exercise and one role play (Vermeiren et al., 2009, p. 308).

In the second half of 2010, ESF decided that test centers were to accredit their exams. Prerequisites for further funding in 2011 were not only validation but also the development of a reliable, objective and valid assessment procedure conform with the *Standard*. As a result, the COC decided to develop a new certification exam.

4. The social interpreter standard

An SI converts oral messages in a social context from a source language into a target language in order to enable efficient communication between the participating parties. He interprets in a complete, neutral and reliable manner. This is the definition of the *Standard* (SERV, 2008).

An SI masters four indispensable competencies: (1) *Processing spoken messages*, (2) *Reproducing spoken messages*, (3) *Deontological conduct*, and (4) *Dealing with deontological conflicts*.

Each competency is made up of several success criteria. The success criteria are the operationalization of the competencies in observable behavior. The competencies *Processing* and *Reproducing spoken messages* also indicate a required knowledge of the B2-level of the Common European Framework of Reference for Languages (CEFR) (See Table 1).

In addition to the competencies, the *Standard* also specifies restrictive instructions for the assessment procedure. The procedure should contain at least one role play and a Dutch and FL proficiency test. The duration of the entire assessment procedure may not exceed 90 minutes.

4 The European Social Fund (ESF) reinforces the implementation and innovation of the Flemish employment policy.

Apart from the *Standard*, ESF also formulates assessment rules. Primarily, all success criteria should be tested twice by means of two separate test moments.

Table 1. The social interpreter standard

1. Processing spoken messages: *Success criteria* 1.1 Understands both the vocabulary and the idea behind the message 1.2 Indicates either verbally or non-verbally that the speaker should soon stop or continue the conversation 1.3 Requests clarification from the speaker when something is not understood or heard 1.4 Makes use of a note-taking technique adapted to one's personal needs in order to record the content in full *Required knowledge* The B2-level of the CEFR for understanding the Dutch and foreign language
2. Reproducing spoken messages *Success criteria* 2.1 Has clear articulation 2.2 Speaks at a sufficiently loud volume 2.3 Maintains an acceptable speech rate when switching between 2 languages 2.4 Asks the speaker for clarification when the translation of a term is either not known or the correctness of a translation or paraphrase is not certain 2.5 Remains as close as possible to the original message in his choice of words or conveys the idea behind the message without losing its meaning 2.6 Converts language-specific expressions and constructions in the source language into expressions and constructions which come across as natural and correct to the users of the target language *Required knowledge* The B2-level of the CEFR for speaking the Dutch and foreign language
3. Deontological conduct *Success criteria* 3.1 Introduces him/herself as an interpreter to both parties at the beginning of the conversation and explains the deontological principles in a way that is adapted to the level of the interlocutor 3.2 Positions him/herself spatially as neutrally as possible and if possible in a triad arrangement 3.3 Avoids private conversations with the client and/or user before, during and/or after the conversation 3.4 Interprets in the first person to foster communication between the client and the user 3.5 Encourages direct eye-contact between the user and the client by pointing this out to both parties 3.6 Translates everything without adding, omitting or changing anything 3.7 Avoids either verbal or non-verbal expressions of personal opinions, preferences, interpretations or feelings

4.	Dealing with deontological conflicts
Success criteria	
4.1	Remains forthcoming and polite whatever may take place
4.2	Repeats the deontological principles whenever something is asked which contradicts the deontological rules
4.3	Immediately notifies both parties when and why an assignment cannot be optimally carried out
4.4	Keeps all unavoidable conversations with the client and/or user prior to, during and/or after the interpreter assignment as neutral as possible
4.5	Gives both parties feedback whenever private conversations could not be avoided in order to safeguard trust and transparency
4.6	Asks the client or user to verbally summarize or go over documents that need to be translated so this can be interpreted
4.7	Limits him/herself to the interpreter assignment and does not assist with other tasks

To pass a competency, the candidate has to master all its success criteria. Secondly, all competencies should be tested at least once via a realistic simulation of the profession. During this simulation, graders have to evaluate the performance through observation. Thirdly, each competency has to be evaluated simultaneously by two independent graders. Ultimately, the exam should approach the reality of the profession as much as possible.

5. The development of the certification exam

Considering all above requirements, the question the COC needed to tackle was how to develop a reliable, objective and valid assessment procedure. The development procedure was split up in several phases.

5.1 Phase 1: the analysis

The first phase consisted of three brainstorming days in which the entire COC-team participated. The first day, the following questions were answered for each success criterion:

(1) Is it possible to test the criterion by means of observation? How can we make it clearly observable during the test? Are there different techniques to do this?
(2) Is the criterion considered as an indispensable criterion to pass the entire competency? Or is it rather regarded as an accessory criterion?
(3) What profile should the grader have in order to be able to test the success criterion?

(4) Is it possible to test all competencies separately? Or are they inextricably connected[5]?

The exercise led to an extensive list of possible techniques to test all criteria. It also created the awareness that all success criteria can be elicited and thus be made observable, but that it depends on the candidate's performance whether he actually executes the elicited behavior or not. Whether graders effectively observe the elicited behavior or not, does not automatically imply a 'pass' or 'fail' success criterion. This is for example the case with success criterion *1.3. Requests clarification from the speaker when something is not understood or heard* (see Table 1). If one wishes to test this criterion during for instance a role play, one can ask the role players to murmur a word during the conversation. In this case, the test elicits the candidate's behavior to ask for clarification because something was not understood or heard. If the candidate does this, he passes the criterion. However, the possibility always exists that the candidate did understand or hear the word and reproduces it correctly in the target language. In this case, asking for clarification is redundant and the graders cannot observe the expected behavior. Thus, non-observation implies a 'pass' criterion. If the candidate did not understand or hear the word and does not ask for clarification leading to a false reproduction of the word in the target language, the candidate fails the criterion.

This leads to three evaluative categories for each success criterion: 'pass', 'fail' and 'not-observed'. Depending on the success criterion the category 'fail' or 'pass' can converge with the category 'not-observed'. This is for example the case with success criterion *3.1. Introduces him/herself as an interpreter to both parties at the beginning of the conversation and explains the deontological principles in a way that is adapted to the level of the interlocutor.* A candidate who does not introduce him/herself at all at the beginning of a conversation, automatically fails this success criterion. In this case the category 'not-observed' converges with the category 'fail'.

Additionally, a fourth category can be implemented, i.e. 'uncertain'. This category implies that a grader is not certain whether the candidate should pass or fail the success criterion. This can be for instance the case with success criterion *3.4. Interprets in the first person to foster communication between the client and the user.* A candidate who constantly switches between using the first and the third person, displays inconsistent behavior. His behavior thus switches between 'pass' and 'fail'. In this case, the category uncertain can be

5 This fourth question was necessary to explore the possibilities of exemption from parts of the exam in case of re-examination. An ESF-rule states that exemption is only possible on the level of competencies and not success criteria.

used, creating space for the deliberation of a success criterion. Consequently, if a candidate passes all success criteria of a competency except one and scores uncertain for this one, he can still pass the entire competency.

Ultimately four categories can be used to evaluate the success criteria: 'fail', 'pass', 'not-observed' and 'uncertain'. All competencies are inextricably connected and can only be tested by means of role plays. An exemption for a certain competency becomes thus impossible.

On the second brainstorming day, the dissection of the CEFR B2-level took place[6]. The description of the B2-level mentions three main categories: *Understanding* (with subcategories *Listening* and *Reading*), *Speaking* (with subcategories *Production* and *Interaction*) and *Writing*. Considering the actual tasks of an SI and the requirements from the *Standard*, the categories *Writing* and *Reading* were eliminated from our analysis. The *Standard* stipulates that written knowledge of the B2-level is not expected from a candidate. The following subcategories thus remained:

(1) *Listening*: I can understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar. I can understand most TV news and current affairs programs. I can understand the majority of films in standard dialect;
(2) *Production*: I can present clear, detailed descriptions on a wide range of subjects related to my field of interest. I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options; and
(3) *Interaction*: I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible. I can take an active part in discussion in familiar contexts, accounting for and sustaining my views.

The *Standard* also mentions that the duration of a language proficiency test should not exceed 15 minutes. Since it would not be feasible to test all above-mentioned criteria within 15 minutes, we excluded even more criteria from our assessment. (1) *I can understand most TV news and current affairs programs. I canunderstand the majority of films in standard dialect*. It would be impossible to test these criteria within 15 minutes. This elimination was pragmatic. (2) *I can take an active part in discussion in familiar contexts, accounting for and sustaining my views*. An SI may never take part in discussions nor account for or sustain his views. His task is to remain as neutral as possible. If we test this criterion, we would expect something of an SI he is not allowed to do in reality. This would make the exam less realistic.

Ultimately, the following criteria remained:

6 The description of the self-assessment grid was used. See also http://www.coe.int/t/DG4/Portfolio/?M=/main_pages/levels.html

(1) *Listening*: I can understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar;
(2) *Production*: I can present clear, detailed descriptions on a wide range of subjects related to my field of interest. I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options; and
(3) *Interaction*: I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible.

The relations between the elements of the remaining criteria were also determined. When it comes to the criterion *Listening*, we have an OR-AND relation. A candidate should either understand extended speech OR a lecture, (AND) in which complex lines of arguments should be followed provided the subject is reasonably familiar. Regarding *Production*, a candidate should either present, detailed descriptions on a wide range of subjects related to his field of interest OR explain a viewpoint on a topical issue giving the advantages and disadvantages of various options. In this case, there is an OR-OR relation. These relations are crucial to make the criteria observable during the exam. Apart from this, we listed the possible ways to test the remaining criteria.

We also agreed that we would test the required B2-knowledge not only during the language proficiency tests but also during the evaluation of the competencies *Processing* and *Reproducing spoken messages*.

Brainstorming day 3 was devoted to designing a prototype of the certification exam taking into account all the decisions of the previous brainstorming days. Simultaneously, the input of the social interpreting agencies and graders was processed. This input had been gathered before the start of the development procedure. Initially this had happened during a Master Class held in 2009 during which the old version of the certification exam was scrutinized in order to improve it. Secondly, they had been given the opportunity to suggest ideas and propose improvements by means of a questionnaire. Taking the input of the agencies into account, was a vital step of the development procedure. The SI certificate has no societal value if the interpreter agencies do not acknowledge the way a candidate obtains this certificate. Additionally, evaluations of the old version of the certification exam by SI candidates were also processed and taken into account.

The developed prototype was presented to the SI agencies and a number of academics from the five Flemish interpreter colleges. These academics had all participated frequently as graders in exam juries. Their input was deemed equally important. The interpreter colleges are the academic partners of the COC. The majority of the graders come from these colleges.

5.2 Phase 2: the experiment

After processing the input of the SI agencies and academics, test material was developed to test the exam. At this point the format of the exam was made up of four evaluative parts: a Dutch and FL proficiency test, Role play 1 and Role play 2. The experiment had several objectives:

(1) Do two independent juries evaluate a candidate's performance in the same manner? Is the exam inter-rater reliable and objective?
(2) Is the exam valid? Does it test what it should test?
(3) Is the exam authentic? Does it approach the reality as much as possible?
(4) Can we test everything within the 90 minutes mentioned in the *Standard*?
(5) How do the graders experience the new format of the exam? Do they have suggestions to improve the test material? and
(6) How do the candidates experience the exam? Do they have certain suggestions?

The experiment was carried out as follows: three different types of candidates were selected to appear for two independent juries. The first type had been certified in the past by means of the older version of the exam. The second had previously taken the older version of the exam, but had not passed it. The third type had never taken a certification exam.

The experiment took place for two different language combinations: Dutch-Spanish and Dutch-Farsi. This was essential because Spanish is listed as a common FL and Farsi is not. Only for common FLs, the *Standard* demands that the required FL knowledge has to be evaluated simultaneously by two independent graders[7]. This implies that juries for common languages always consist of an extra grader in comparison to juries for uncommon FLs.

Each candidate took the exam in the physical presence of one jury while the entire performance was videotaped. The first jury evaluated the performance immediately. Later, a second jury, unaware of the first jury's decision, evaluated the candidate's performance while watching the videotaped performance.

The results of the inter-rater reliability experiment were very positive. The results of both juries corresponded entirely for all the language proficiency tests and all the role plays on the level of the competencies. However, there were minor differences on the level of some success criteria. The latter was due to minor differences in the interpretation of these success criteria. In advance, the four

7 The common FLs listed in the *Standard* are: Bosnian-Serbo-Croatian, Danish, German, English, French, Italian, Polish, Portuguese, Russian, Spanish, Standard Arabic, Standard Chinese and Turkish.

evaluative categories of all success criteria had been meticulously described in the *Appropriate-Inappropriate Behavior Guide.* This *Guide* had been written in order to limit the interpretations of different graders and as a reference point in case of discussion between graders. Despite its availability, interpretation differences had nevertheless arisen. Ultimately, this led to a slight adjustment of the description of the success criteria for which differences had occurred.

5.3 Phase 3: the finalization

The final phase consisted of processing the results of the experiment. The format of the exam had proved to be valid, reliable and authentic, but did exceed the 90 minutes. This was largely because the role plays were simply too long. Therefore, the length of the role plays was shortened. The feedback of the graders and candidates was also processed, resulting in some inconsequential adjustments of the procedure and evaluation grids.

The final exam format is made up of four evaluative parts, two pauses for the candidate and three evaluative moments for the graders. All candidates appear before a jury. The chairperson ensures all procedures are carried out correctly and evaluates the candidate's Dutch and deontological conduct. Since all competencies have to be observed by two graders independently, there is a second grader for Dutch and deontological conduct. Depending on the FL, there are one or two FL graders. FL graders are by no means allowed to evaluate the Dutch language or deontological conduct. The chairperson and extra Dutch grader are not allowed (and usually not able) to evaluate the candidate's FL.

The exam procedure is composed as follows:

(1) An ice-breaker (5 min): This part comprises of an introduction of the jury, the explanation of the exam procedure and a short conversation with the candidate in Dutch. This conversation is not evaluated. Its purpose is to put the candidate at ease.

(2) A Dutch and FL proficiency test (10 min each): The language proficiency test measures whether the candidate has the B2-level for understanding and speaking. The candidate can choose whether he wishes to start with his strongest language or not. Candidates believe this option reduces their stress. Each test consists of two parts.

(2a) A 3-minute *listening exercise* with six questions tests the *Listening*, *Production* and *Interaction* criteria. The candidate may not take notes during the listening exercise. The audio clip is played only once. Three questions simply require a YES/NO-answer. These questions test the understanding skills. A correct answer gives the candidate 1 point. The other three questions are open. These

test the understanding and production skills. The open questions are evaluated on two levels. If the answer is correct concerning the content, the candidate receives 1 point. If it is also grammatically correct, the candidate receives a second point. All six questions test the interaction skills since the questions are posed orally by a grader. If a question is not understood or heard clearly, the candidate can ask for a repetition only once. A score of six out of nine (6/9) is required to pass the listening exercise. This part takes about 7 minutes.

(2b) During the second part, the candidate has to *describe a set of images* as detailed as possible. This exercise tests spontaneous production and interaction. The candidate has to speak for 3 minutes. Graders can ask additional standard questions to elicit more production of the candidate. Graders evaluate the production skills by means of three subcategories: pronunciation, grammar and vocabulary. A candidate passes this second part of the proficiency test if he passes each subcategory.

A candidate passes the entire proficiency test when he passes the two parts. Deliberation of the listening exercise is possible if he does not have the required 6/9 score for the listening exercise. A minimum score of 3/6 is necessary for the questions that test the candidate's understanding, i.e. YES/NO-questions and open questions on the level of the content. So if a candidate did not pass the listening exercise due to weak production, and not due to incorrect understanding, than he can still demonstrate a good production level during the description of the images. This procedure creates an equal balance between the evaluation of the criteria *Understanding* and *Production*.

(3) Pause/Evaluation 1 (10 min): During this pause, the FL grader(s) evaluate(s) the FL proficiency test, while the graders for Dutch evaluate the Dutch proficiency test. Initially, graders evaluate independently. The chairperson brings the separate evaluations together. In case of correspondence between the different graders, there is no problem. In case of different opinions the chairperson moderates the discussion until a consensus is found. This pause was inserted for two different reasons. Firstly, the pause guarantees that all parts of the exam are evaluated separately. Since the B2-level is evaluated again during the role plays, the jury wants to determine whether the candidate has a B2-level freestanding from an interpreting context. Secondly, the candidate can use the pause to prepare the next part of the exam.

(4) Role play 1 (average 15 min): A role play is a simulation of a real social interpreting conversation. The exam format consists of two role plays because all competencies should be tested twice separately and it is impossible to elicit all success criteria in one role play. The latter would make a conversation highly unrealistic. Role play 1 is the first test moment for all the competencies, success criteria and the language proficiency during a role play. The *Standard* mentions

that a role play lasts minimally 10 and maximally 20 minutes. Its role players should approach the level of native speakers. The contexts of Public Center for Social Welfare, Center for Welfare Work and psychosocial assistance, residence status and refugee-related subjects, integration, education, health and its social aspects and child & family have to be used. A role play may contain a maximum of six deontological conflict situations of which maximum three are provoked by the user and maximum three by the client. A role play may contain a maximum of 35 vocabulary difficulties of which maximum 15 regarding general vocabulary, maximum 15 regarding terminology and expressions related to the social context, maximum 3 acronyms related to the social context and maximum 2 idiomatic expressions.

It was decided that Role play 1 should focus more on vocabulary difficulties than on deontological conflict situations. A week before the exam, the candidate receives the context, topic and interlocutors of Role play 1. This correlates with interpreting in face-to-face situations whereby an SI receives this information from the SI agency a few days before the assignment takes place. One can thus prepare for an assignment. Similarly, a candidate – while studying for the exam – can look up vocabulary regarding the topic of Role play 1.

Considering the above, all standardized scripts of Role play 1 contain 35 vocabulary difficulties, 2 deontological conflicts and take averagely about 15 minutes. If the candidate's performance exceeds 20 minutes, he can no longer pass Role play 1. It is an indication that he does not interpret smoothly enough (success criterion 2.3). Apart from this, an SI should be able to interpret consecutively up to 3 minutes. Although this is not explicitly mentioned in the *Standard*, it is a criterion which the entire sector deems utterly important. Interpreting consecutively enhances the trust relation between the user and client. A piece of text of about 2 minutes acted out by the Dutch user is inserted in all the scripts of Role play 1. The candidate has to be able to reproduce this longer consecutive part completely and accurately in the FL.

(5) Pause/Evaluation 2 (10 min): This pause is used to evaluate the four competencies and the candidate's language proficiency during Role play 1. The chairperson and the grader for Dutch and deontological conduct evaluate all four competencies and ascertain whether the Dutch proficiency meets the CEFR B2-level. The FL grader(s) only evaluate(s) the competencies *Processing and Reproducing spoken messages* and ascertain(s) whether the FL proficiency meets the CEFR B2-level. The chairperson and Dutch grader evaluate the competencies *Processing and Reproducing spoken messages* on the level of the Dutch language, while the FL graders do this on the level of the FL. Two different standardized evaluation grids are used for Role play 1. One is used while the role play takes place. The grader can follow the script, indicate mistakes or comment on the deontologi-

cal conduct. A second grid sums up all the success criteria and the four evaluative categories ('fail', 'pass', 'not-observed' and 'uncertain'). This grid is actually a summary of the candidate's behavior. Firstly, each grader completes this summary based on his observation of the role play. Then the chairperson brings all evaluations together. A candidate needs to pass all success criteria for all graders in order to pass Role play 1.

This pause/evaluation was inserted for similar reasons as Pause/Evaluation 1. Graders wish to evaluate the candidate's performance independently from the candidate's performance during Role play 2. If one would wait till the end of the entire exam to evaluate Role play 1, a grader's evaluation might become distorted by what he has observed during Role play 2. So this evaluation ensures more objectivity. Additionally, this pause gives the candidate the possibility to prepare Role play 2.

(6) Role play 2 (average 15 min): Role play 2 is the second test opportunity for all competencies, success criteria and the language proficiency during a role play. There are some significant differences between Role play 1 and 2. Firstly, the candidate is not informed a week in advance about the topic of the role play. He receives this information right before Pause 2. This correlates with interpreting provided over the phone whereby an SI receives an interpreting assignment of the SI agency right away. Thus, one cannot prepare for an assignment. During the exam, the candidate can use the 10 minute pause to look up some information. Since the candidate had less preparation possibility, Role play 2 focuses more on deontological conflicts than on vocabulary difficulties. It contains a maximum of 35 vocabulary difficulties, four deontological conflicts and takes averagely about 15 minutes. Role play 2 also contains a longer consecutive part. The FL client will act out a piece of text lasting about 2 minutes, that has to be reproduced by the candidate in Dutch. So the language direction is reversed in comparison to Role play 1. The context in which Role play 2 takes place will never be the same as Role play 1. This is to determine whether the candidate masters at least two subjects of the wide range of subjects SIs usually deal with.

(7) Final Evaluation (15 min): The graders evaluate Role play 2 in the same manner as Role play 1. They also formulate a final evaluation about the candidate's performance. This feedback is sent to the candidate within one week. A candidate passes the entire exam, if he passes all four parts of the exam. In case of failing, a candidate can retake the exam and will receive advice on how to remedy his shortcomings. He can take two exams per year. If the candidate previously passed a language proficiency test, he is exempted from this part of the exam. Exemption from the role plays is only possible if the candidate passed Role play 1 and 2. The candidate needs to prove that he can pass the two test moments for the competencies during the same exam.

6. Fair assessment measures

Equality is one of the two ethical principles of the SI certification exam. All candidates – regardless of their degree, experience or (interpreter) language – must pass the same exam in order to gain access to the SI labor market. Exemption from the exam is not possible. Candidates are allowed to follow courses or have preparatory meetings with a consultant in order to prepare for the exam. Equity is the second principle. The exam and SI certificate offer chances to immigrants who are generally disadvantaged on the labor market due to the absence of (recognized) degrees. The SI certificate grants them opportunities to improve their professional status. The exam is free of charge and adapted to the visually handicapped.

To ensure equality and equity, maximizing objectivity is a central concern of the exam procedure. It is done via applying a number of quality control mechanisms.

6.1 Equalizing test materials

Any candidate – regardless of the language combination in which he is tested – should be subjected to the same levels of difficulties. How can one ensure this, when test developers do not master all the languages being tested?

The texts used for the listening exercises are selected according to a rigid selection procedure. Primarily, all texts should adhere to a number of criteria. They should be based on a news item (radio or newspaper) originally spoken or written in the concerning language. The length of the text is between 300 and 400 words and the spoken version should last about 3 minutes. The subject of the test should be relevant to an SI for instance health care, asylum seekers, migrants, social security, labor, education, minorities, integration, racism, discrimination and psychosocial care. The nature of the text is preferably narrative, for example an immigrant explains how he experienced his integration process.

The selection of Dutch language texts and the formulation of the questions is always done by a COC employee. A second COC employee controls the abovementioned text criteria and the difficulty level of the questions. If necessary, the text or questions are adjusted or rejected. Since the COC employees do not master all FLs tested during the exam, a different selection procedure is in place for the FL listening exercise. An FL grader is requested to select a text and translate it to Dutch. A COC employee decides based on the translation whether the text complies to the criteria and formulates six questions in Dutch. The latter makes sure

that all questions for all languages are equally difficult. A second COC employee checks the Dutch translation and questions. If all is approved, the FL grader translates the questions to the FL.

All texts are recorded at a similar speaking rate by a grader for Dutch or the relevant FL grader. Standardized evaluation grids are used by all graders to evaluate the listening exercise. They always contain the text, the questions, the correct answers and extra space to write comments about the candidate's grammar, articulation and vocabulary.

Congruency of the difficulty level of the role plays is mainly ensured by the fact that all scripts of Role play 1 and 2 should conform to the *Standard* criteria. All role plays are based on real social interpreting conversations and written by COC employees. Apart from the criteria of the *Standard*, the COC formulated extra criteria. The total number of words of Role play 1 should be between 1000 and 1070 and those of Role play 2 between 900 and 950. The longer consecutive piece inserted in both role plays should count between 230 and 300 words. All role plays are entirely written in Dutch. The parts of the FL client are translated by the FL role player.

The use of standardized evaluation grids congruent with the criteria being evaluated, enhances the grader's objectivity and limits possible interpretations. Each part of the exam has relevant evaluation grids. Also the availability of a detailed procedural manual for each grader and the *Appropriate-Inappropriate Behavior Guide* contribute to this aim.

6.2 Graders

All candidates appear for parallel composed juries. Detailed profiles have been written for each grader and his corresponding tasks. The chairperson is always a COC employee. He is an SI expert and has been hired as a COC employee based on his competencies.

The grader for Dutch and deontological conduct is usually connected to an interpreting college. He should have experience as a Dutch language assessor and an extensive knowledge of an SI's deontological code and interpreter techniques.

An FL grader does usually not work for the COC either. He needs to have a sufficient level of Dutch to participate in the jury. His command of the FL approaches the level of a native speaker. He needs to use the FL actively on a systematic basis. Preferably he is connected to an interpreter college, university or vocational school where he teaches the language. The latter implies that FL graders are reasonably easy found for all the common FLs tested, but it regularly poses great difficulties for some of the uncommon. In this case, there are some

preferences for the selection of FL graders: he has a degree in linguistics or interpreting/translation, is a journalist or author in the FL and has some knowledge of the Flemish social sector. An FL grader may never be an active SI in Flanders to avoid possible conflicts of interests.

There is a strict selection procedure for graders who are not COC employees. All graders are screened on the basis of their curriculum vitae and interviewed by two COC employees. Due to limiting financial constraints and practical hindrances, there is no possibility to screen the command of the FL of the grader[8]. The command of the Dutch language is not officially screened but is a focal point during the interview.

Once selected, all graders receive a compulsory training before they appear in a jury. Usually this is a one-time training, but if requested or needed, more training can be provided. All graders should prepare properly by reading the procedural manual, the *Appropriate-Inappropriate Behavior Guide* and the role plays. On exam days the chairperson provides a 30 minute personal training during which the exam principles and the use of the evaluation grids are explained elaborately. When finally evaluating, a grader does never function in a vacuum. Grading is a part of teamwork under the supervision of the chairperson. The latter functions as the reference in case of a grader's doubt about an element of the procedure. The continuous monitoring of the adherence to the exam procedure by the chairperson constitutes another important mechanism to ensure fairness.

Another element contributing to fairness, is the grader's code of ethics. All graders need to endorse this code. Important elements are the assurance of neutrality and impartiality, not divulging exam material or test results and the protection of the candidate's privacy.

6.3 Role players

Each role play requires two role players, i.e. a Dutch and an FL speaking one. All role players receive a short training. They observe an experienced role player, learn a number of techniques on how to deal with unforeseen circumstances and in which cases improvisation is allowed.

The Dutch role players are usually native speakers who work for the mother organization *Junction Migration-Integration*. The FL player in a jury for common FLs, is always one of the two FL graders of the jury. In case of a jury with one FL grader, a certified SI – if one exists and is available – is employed to play the

8 For some FLs we do not even find a suitable grader, so finding an assessor to grade the (future) grader is often impossible.

role of the client. It is a very strenuous task to act as a role player and a grader simultaneously. This does not imply that the FL role player in a jury for common languages does no longer have to evaluate the candidate's performance. However, it makes the task less strenuous because there is still a second FL grader who – in case of doubt – can serve post-factum as a reference.

Role players also have to endorse a code of ethics mentioning the same principles as the graders' code of ethics. It is clearly mentioned that role players who are not graders (i.e. Dutch or the certified SI), should not assess or pass any judgments about the candidate's performance.

6.4 Complaint commission

A candidate who does not agree with his results or has doubts about a proceeding of his certification exam, may appeal to the complaint commission. This appeal is free of charge. Firstly a candidate is invited to listen and watch his taped exam performance and discuss his arguments. When the candidate still disagrees, he should write a clearly motivated appeal letter to the COC. A commission comprising one representative of the COC, one of an SI agency and one of an interpreting college, will review the exam. If the commission wishes to consult an FL expert or the candidate, an interview may take place. The commission will form a new evaluation. This decision is final.

7. The accreditation procedure of the European Social Fund

The European Social Fund requires a test center to employ a valid and reliable test in order to receive funding. In 2010, the COC test format went through an accreditation process. Preparation of this process included all the above-mentioned steps and completing a very detailed ESF-template with standard questions about the exam procedure. Firstly, the test was evaluated by peer test centers. Secondly, all test material including the ESF template were sent to a number of (unknown) experts selected by ESF. Evaluation criteria were validity, reliability and conformity with the *Standard*.

The COC exam format was evaluated positively. The consequences of this accreditation are significant. If another institution in Flanders wants to become an SI test center, the COC exam format must be employed. Additionally, the COC was granted funds to organize 320 exams in 2011 and 2012.

In 2011, 161 exams took place. The language combinations tested were Dutch-Albanian, Arabic, Armenian, Bosnian-Serbo-Croatian, Bulgarian, Chinese,

Czech, English, Farsi, French, German, Hungarian, Italian, Lithuanian, Mongolian, Polish, Portuguese, Romanian, Russian, Somali, Spanish, Tamazight and Turkish. The passing rate was 27%.

8. The future

Despite this accreditation and the Flemish Integration Decree stimulating the use of qualitative SIs, the future of the COC as a certifying organization or test center is insecure.

There are presently two currents potentially endangering the entire system. The SERV wishes to reform the system of certifying competencies related to a specific profession. An SERV research has proved that the acquisition of a title of professional competency does usually not guarantee a job or one's improvement on the labor market. The mere fact that one has a hairdresser certificate, for example, does not exempt him/her from further selection processes of the salon were he wishes to work. Also employers seem largely unaware of the value of these certificates. Consequently, the SERV wishes to create a system to test competencies independently and not connected to a certain profession. This SERV argument does clearly not count for the SI sector. He who holds an SI certificate, will always receive SI assignments from a social interpreting agency. There are no further selection processes to work for a social interpreting agency. This proves the societal value of the SI certificate which other professional certificates apparently lack.While the SERV has assured that the current existing test centers will not simply be abolished, ESF has made it clear it will no longer provide finances. The Flemish Ministry of Integration has not yet shown any sing to structurally embed finances for SI certification in the government budget.

A second potential danger is the current political climate. As in many Western European countries, right-wing parties win more and more ground in Flanders and Belgium.The government is increasingly questioning the necessity of SIs as such. Immigrants for which they are employed, should be stimulated to learn Dutch and to deal with social welfare workers without the use of SIs. While some politicians understand that SIs mainly facilitate the working conditions of the social welfare worker, they still think a time limit should be installed on how long an immigrant is entitled to an SI after his arrival in Belgium. They usually forget that it is not the immigrant who requests the presence of an SI, but the social welfare worker. Others do still not recognize the quality a certified SI brings whose competencies have been confirmed, and wish to install a system of SI volunteers. Additionally, the Flemish government decided in July 2011 to reform the entire integration and immigration sector of which SIs are a part. At this point it is nei-

ther clear what consequences this could have on the certification of SIs, nor on their existence at all.

References

Centrale Ondersteuningscel voor Sociaal Tolken en Vertalen (COC). 2007. *Sectoraal dossier sociaal tolken en vertalen*. Brussels.

Sociaal-Economische Raad van Vlaanderen (SERV). 2008. *Standaard Ervaringsbewijs sociaal tolk (m/v).*Brussels: SERV. See also: http://www.ervaringsbewijs.be/beroepen/documenten/Standaard_sociaal_tolk.pdf [accessed 2012.01.16].

Sociaal-Economische Raad van Vlaanderen (SERV) and Centrale Ondersteuningscel voor Sociaal Tolken en Vertalen (COC). 2007. *Beroepscompetentieprofiel Sociaal Tolk*. Brussels: SERV. See also: http://www.serv.be/sites/default/files/documenten/pdfpublicaties/1246.pdf [accessed 2012.01.16].

Vermeiren, H., Van Gucht, J. & De Bontridder, L. (2009). Standards as critical success factors in assessment. Certifying social interpreters in Flanders, Belgium. In Angelelli, C.V. & Jacobson, H.E. (Eds.) *Testing and Assessment in Translation and Interpreting Studies* (pp. 297–329). Philadelphia: John Benjamins Publishing Company.

Vlaamse Gemeenschap. 2004. *Decreet betreffende het verwerven van een titel van beroepsbekwaamheid*. Brussels: Belgisch Staatsblad. See also: http://www.ejustice.just.fgov.be/mopdf/2004/11/26_2.pdf#Page22 [accessed 2012.01.16].

Vlaams Parlement. 2009. *Decreet tot wijziging van het decreet van 28 april 1998 inzake het Vlaams beleid ten aanzien van etnisch-culturele minderheden*. Brussels: Belgisch Staatsblad. See also: http://docs.vlaamsparlement.be/docs/stukken/2008–2009/g2154-4.pdf [accessed 2012.01.16].

Language Testing and Evaluation

Series editors: Rüdiger Grotjahn and Günther Sigott

Vol. 1 Günther Sigott: Towards Identifying the C-Test Construct. 2004.

Vol. 2 Carsten Röver. Testing ESL Pragmatics. Development and Validation of a Web-Based Assessment Battery. 2005.

Vol. 3 Tom Lumley: Assessing Second Language Writing. The Rater's Perspective. 2005.

Vol. 4 Annie Brown: Interviewer Variability in Oral Proficiency Interviews. 2005.

Vol. 5 Jianda Liu: Measuring Interlanguage Pragmatic Knowledge of EFL Learners. 2006.

Vol. 6 Rüdiger Grotjahn (Hrsg. / ed.): Der C-Test: Theorie, Empirie, Anwendungen/The C-Test: Theory, Empirical Research, Applications. 2006.

Vol. 7 Vivien Berry: Personality Differences and Oral Test Performance. 2007.

Vol. 8 John O'Dwyer: Formative Evaluation for Organisational Learning. A Case Study of the Management of a Process of Curriculum Development. 2008.

Vol. 9 Aek Phakiti: Strategic Competence and EFL Reading Test Performance. A Structural Equation Modeling Approach. 2007.

Vol. 10 Gábor Szabó: Applying Item Response Theory in Language Test Item Bank Building. 2008.

Vol. 11 John M. Norris: Validity Evaluation in Language Assessment. 2008.

Vol. 12 Barry O'Sullivan: Modelling Performance in Tests of Spoken Language. 2008.

Vol. 13 Annie Brown / Kathryn Hill (eds.): Tasks and Criteria in Performance Assessment. Proceedings of the 28th Language Testing Research Colloquium. 2009.

Vol. 14 Ildikó Csépes: Measuring Oral Proficiency through Paired-Task Performance. 2009.

Vol. 15 Dina Tsagari: The Complexity of Test Washback. An Empirical Study. 2009.

Vol. 16 Spiros Papageorgiou: Setting Performance Standards in Europe. The Judges' Contribution to Relating Language Examinations to the Common European Framework of Reference. 2009.

Vol. 17 Ute Knoch: Diagnostic Writing Assessment. The Development and Validation of a Rating Scale. 2009.

Vol. 18 Rüdiger Grotjahn (Hrsg. / ed.): Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from Current Research. 2010.

Vol. 19 Fred Dervin / Eija Suomela-Salmi (eds. / éds): New Approaches to Assessing Language and (Inter-)Cultural Competences in Higher Education / Nouvelles approches de l'évaluation des compétences langagières et (inter-)culturelles dans l'enseignement supérieur. 2010.

Vol. 20 Ana Maria Ducasse: Interaction in Paired Oral Proficiency Assessment in Spanish. Rater and Candidate Input into Evidence Based Scale Development and Construct Definition. 2010.

Vol. 21 Luke Harding: Accent and Listening Assessment. A Validation Study of the Use of Speakers with L2 Accents on an Academic English Listening Test. 2011.

Vol. 22 Thomas Eckes: Introduction to Many-Facet Rasch Measurement. Analyzing and Evaluating Rater-Mediated Assessments. 2011.

Vol. 23 Gabriele Kecker: Validierung von Sprachprüfungen. Die Zuordnung des TestDaF zum Gemeinsamen europäischen Referenzrahmen für Sprachen. 2011.

Vol. 24 Lyn May: Interaction in a Paired Speaking Test. The Rater´s Perspective. 2011.

Vol. 25 Dina Tsagari / Ildikó Csépes (eds.): Classroom-Based Language Assessment. 2011.

Vol. 26 Dina Tsagari / Ildikó Csépes (eds.): Collaboration in Language Testing and Assessment. 2012.

Vol. 27 Kathryn Hill: Classroom-Based Assessment in the School Foreign Language Classroom. 2012.

Vol. 28 Dina Tsagari / Salomi Papadima-Sophocleous / Sophie Ioannou-Georgiou (eds.): International Experiences in Language Testing and Assessment. Selected Papers in Memory of Pavlos Pavlou. 2013.

Vol. 29 Dina Tsagari / Roelof van Deemter (eds.): Assessment Issues in Language Translation and Interpreting. 2013.

www.peterlang.de