Dina Tsagari / Roelof van Deemter (eds.)

Assessment Issues in Language Translation and Interpreting
Assessment Issues in Language Translation and Interpreting
Language Testing and Evaluation
Series editors: Rüdiger Grotjahn and Günther Sigott

Volume 29
Assessment Issues in Language Translation and Interpreting
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
San Diego University

Translation and Interpreting in Language 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 Cyprus. 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


Part I
General Issues in Assessing Language Translation and Interpreting
How Do We Assess Students in the Interpreting Examinations?

Fred S. Wu
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-
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

<table>
<thead>
<tr>
<th>Student / Code (pseudonyms)</th>
<th>Course exam mark</th>
<th>A Language</th>
<th>B Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ally / A</td>
<td>low 50</td>
<td>Chinese</td>
<td>English</td>
</tr>
<tr>
<td>Beth / B</td>
<td>mid 60</td>
<td>Chinese</td>
<td>English</td>
</tr>
<tr>
<td>Cherry / C</td>
<td>70+</td>
<td>English</td>
<td>Chinese</td>
</tr>
<tr>
<td>Daisy / D</td>
<td>mid 50</td>
<td>Chinese</td>
<td>English</td>
</tr>
<tr>
<td>Eileen / E</td>
<td>high 50</td>
<td>Chinese</td>
<td>English</td>
</tr>
</tbody>
</table>

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 pre-selected, 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 \( \frac{n(n-1)}{2} \) 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.

\[
\begin{array}{c|c|c|c|c}
\text{Better} & \text{Worse} \\
\hline
\text{Cherry} & 0.845 & \text{Eileen} & -0.262 & \text{Daisy} & -0.502 & \text{Ally} & -0.755 \\
\end{array}
\]

\(^2\) PC: \( F(4,154) = 37.097, p < 0.001 \); OJ: \( F(4,154) = 42.524, p < 0.001 \); OM: \( F(4,154) = 16.792, p < 0.001 \)
procedures, the examiners were asked to compare the students’ performances in pairs. Given \( n \) students, there should be \( \frac{n(n-1)}{2} \) 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 judgments 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 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

<table>
<thead>
<tr>
<th>Better</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cherry</td>
<td>Beth</td>
</tr>
<tr>
<td>0.845</td>
<td>0.675</td>
</tr>
</tbody>
</table>

Figure 2. OJ Thurstone scales of interpreting proficiency

<table>
<thead>
<tr>
<th>Better</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cherry</td>
<td>Beth</td>
</tr>
<tr>
<td>0.933</td>
<td>0.632</td>
</tr>
</tbody>
</table>

Figure 3. OM Thurstone scales of interpreting proficiency

<table>
<thead>
<tr>
<th>Better</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cherry</td>
<td>Beth</td>
</tr>
<tr>
<td>0.837</td>
<td>0.688</td>
</tr>
</tbody>
</table>

*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

<table>
<thead>
<tr>
<th></th>
<th>Intra-class correlation</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lower bound</td>
</tr>
<tr>
<td>Paired Comparison</td>
<td>Single measures</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>Average measures</td>
<td>0.97</td>
</tr>
<tr>
<td>Overall Judgement</td>
<td>Single measures</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>Average measures</td>
<td>0.97</td>
</tr>
<tr>
<td>Overall Mark</td>
<td>Single measures</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>Average measures</td>
<td>0.95</td>
</tr>
</tbody>
</table>

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
The low values of the three single-measures ICC scores (0.49, 0.52, 0.41) suggest poor and unacceptable consistency levels of the examiners’ judgments. This is indicative of the reliability level when only a single item is used. The OM method has the lowest score (0.41) of Cronbach’s alpha (ICC), which indicates the reliability level when only a single item is used.

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 levels 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

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

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

<table>
<thead>
<tr>
<th>Assessment criteria</th>
<th>Conceptual properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presentation and Delivery</td>
<td>3 aspects:</td>
</tr>
<tr>
<td></td>
<td>• acoustic</td>
</tr>
<tr>
<td></td>
<td>• word/phrase</td>
</tr>
<tr>
<td></td>
<td>• flow of information</td>
</tr>
<tr>
<td>Fidelity and Completeness</td>
<td>3 areas:</td>
</tr>
<tr>
<td></td>
<td>• content accuracy</td>
</tr>
<tr>
<td></td>
<td>• speaker intention</td>
</tr>
<tr>
<td></td>
<td>• contextual consistency</td>
</tr>
<tr>
<td>Audience Point of View</td>
<td>• to have the confidence in the speaker (via the delivery style of interpretation)</td>
</tr>
<tr>
<td></td>
<td>• to receive the speaker’s message at an acceptable level of faithfulness.</td>
</tr>
<tr>
<td>Interpreting Skills and Strategies</td>
<td>• resourcefulness: the ability to use skills and strategies, such as paraphrasing, summarising, skipping, self-correction, background knowledge and anticipation.</td>
</tr>
<tr>
<td></td>
<td>• 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.</td>
</tr>
<tr>
<td>Foundation Abilities for Interpreting</td>
<td>• listening comprehension</td>
</tr>
<tr>
<td></td>
<td>• aptitude and personality</td>
</tr>
</tbody>
</table>

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-
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

![Diagram of the basic conceptual model of interpreting examinations](image-url)