Assessing Simultaneous Interpreting

A study on
Test reliability and
Examiners' assessment behaviour

Shao-Chuan Wu
PhD Thesis

submitted at the School of Modern Languages
Newcastle University

August 2010

To my wife Chen-huei

and

my two children Kevin and Joanne

who accompanied me, and grew up

through the journey

Acknowledgements

I can never thank enough my main supervisor, Dr. Francis Jones, for his patient guidance, insightful advice, and timely encouragement throughout the duration of the study. Without him, this thesis would not have been completed.

I also want to thank my second supervisor Scott Windeatt for meeting with me, giving me ideas, and especially for tricking me into making a conference presentation of my research study in front of a group of esteemed scholars in language testing, whom I had no idea about their identities at that time. It was a wonderful experience and I received many encouraging comments at the conference.

A hearty thank you to Eric Liu, my colleague and mentor at the School of Modern Languages, who shouldered many workloads for me during the long years of this research study. Many thanks also go to all my colleagues who shared interests in the T&I studies, and gave me encouragement and assistance during difficult times.

I also want to thank the following two institutions in Taiwan – Graduate Institute of Translation and Interpretation at National Taiwan Normal University, Taipei; and the Department of English at Wenzao Ursuline College of Languages, Kaohsiung. They very kindly allowed me to use their staff rooms and offices to conduct many of the examination simulations and interviews.

Special thanks must also go to all the participant examiners in this study, who unreservedly offered their time, and shared their professional experiences and insightful comments in the interviews. This research study would not have been such a fruitful and rewarding experience without them.

Last, but certainly not least, I sincerely thank my two external examiners, Professor June Eyckmans and Dr Séverine Hubscher Davidson, for their encouragement and their thorough scrutiny of this thesis. Their incisive comments and helpful suggestions for revision made the thesis so much more readable, though any remaining inadequacy in the presentation and quality of the thesis must be my own responsibility.

Abstract

A substantial amount of research work has been done on the quality assurance of conference interpreting, yielding useful guidelines for the selection and training of interpreters. However, the field of assessment in interpreter training within the educational context is still under-researched. Many interpreter trainers and researchers have pointed out some urgent issues to be addressed in the assessment of interpreting. Among them, the issues surrounding the test validity and reliability are most in need of clarification. This study tackles this complex subject by firstly exploring what the examiners are really paying attention to when assessing student interpreters.

Thirty examiners, who are mostly based in Taiwan – some with substantial interpreting experience and some with less, were invited as study subjects to participate in a simulation of simultaneous interpreting examination. This research study adopted a multi-strategy approach in collecting and analysing research data – quantitative and qualitative. Inconsistencies and fuzziness are two themes identified in the study findings in terms of the examiners' judgements and their use of assessment criteria. The examiners might appear to be using the same assessment criteria, but there were variations in the way how they were used.

This study explores, discusses, and clarifies the intricate relations of various components and factors in the interpreting examinations. Based on the study findings, a conceptual model is proposed as a framework for describing the test constructs of the interpreting examinations, and for understanding how the examiners apply the assessment criteria in order to improve the assessment instruments and examination procedures. At the end, implications of the study method are discussed and suggestions are made for future studies in this area.

Table of contents

List of tables

List of figures

Prelude

It is a week before the mid-term exam for the simultaneous interpreting class. Both the teacher and the students are busy preparing for it. The teacher has informed the students of the subject area of the speech. It is a speech on the topic of how tourism in Asia is recovering from an epidemic scare that crippled the industry. Students set out to work hard preparing for the coming exam.

The exam proceeds as planned. The students come into the exam room one by one, sit in the booth, and interpret the messages into their first language as they listen to the speech. The examiners then discuss each student's performance and give an agreed mark. Depending on the agreed practices within the institution, where there are very different opinions among the examiners, the final mark is given by averaging the marks of each examiner, or moderated by the chairperson of the exam panel.

When the assessment is over and all students have received their marks, they feel that the marks are "fair" and more or less in line with what they have perceived in their class performances; everyone seems happy. Some students may have doubts about their lower-than-expected marks, so they approach the teacher for more detailed feedback. The teacher may need to refer to the notes taken during the exam panel, or to the students' recordings of the exam and the speech script, if there is one, for giving feedbacks. The teacher points out the parts where the students have not done well in the exam, explains the reasons why they haven't got a higher mark (a list of assessment criteria on paper would be useful at this point), and then suggests some practice they can do to improve. The students seem satisfied and go on with their practice.

But the teacher, looking at the marks and speech scripts, seems troubled. Why does this error cost the students more points than that error? Student A has got two major errors but her voice is so pleasant to listen to. Student B has no major error but has

about a dozen small, yet annoying, mistakes and fillers in her delivery. Is it reasonable for Student A to have a higher mark of 64 than Student B's 58, especially when B's performance in class has been consistently good in terms of faithfulness? What does it mean anyway when there's a 6-point difference in their marks? Compared with a dozen small mistakes, does it mean that the two major errors are not significant? In the real world the audience may indeed feel that A is a better interpreter in spite of the two major errors. Can A's pleasant voice really compensate for the two major errors and justify a higher mark? Could the audience in the real world have picked up the mistakes? Have the examiners picked up every error in each student's interpreting performance? If not, was this exam fair? If this exam had been assessed by another panel of examiners, would they have given similar marks to those that the examiners gave in this panel? Moreover, there were some differences in each version of the impromptu speech in the exam. Would this have any effect on students' performances and marks?

And finally, what can be done next time to improve the test design and exam procedure? The interpreter teacher embarks on a journey to find the answers.

CHAPTER 1

The Journey Begins

1.0 An overview of the research background

Interpreting denotes the facilitating of oral communication between different languages. It is "an immediate form of translational activity, performed for the benefit of people who want to engage in communication across barriers of language and culture" (Pöchhacker, 2004: 25). For the communication in many international conferences, the immediate form takes to its highest level to *simultaneous*: interpretation is provided to the audience in real time as the speakers deliver their speeches. So that conference delegates from different cultures and language backgrounds can communicate without delays. Conference interpreters, who work behind the scene in a sound-proof booth, are the core members of a team that facilitate this multilingual communication process. Very often, the quality of the conference interpreters' work affects the quality and efficiency of communication at conferences.

With the ever increasing demand of multilingual communication in the international arena, there has also been a boom in the demand of conference interpreters. The number of educational programmes for interpreters at university level rose from 49 to 80 in the 1960s and 1980s, to 250 in the 1990s, and to an estimation of 300 at the turn of the twenty-first century (Sawyer, 2004: 1). The underlying principle for interpreter education has been the "lasting tradition of training by **apprenticeship**" that was established by the first-generation teachers of interpreting who themselves were accomplished professionals (Pöchhacker, 2004: 175). Nevertheless, with the growing pedagogical influence of language teaching, the traditional apprenticeship approach was questioned for interpreter education in the 1980s. Since then, more rigorous approaches have been suggested in developing curriculum and assessment for training conference interpreters (ibid: 176).

To ensure the quality of interpreting, assessment is necessary in both the professional practice and educational training. Assessment plays an important gate-keeping role to ensure that only suitably qualified interpreters are endorsed to enter the job market, such as the professional examinations in the field of the interpreting profession, and the final examinations at the end of a training course. In the field of professional practice, assessment is a planned activity for the quality assurance of interpreting service; in education, assessment usually refers to making a judgment about students' learning in order to identify their strengths and weaknesses, which usually involves assigning a mark or a grade to their performances and achievements.

It is important to be clear about the role and purpose of assessment in education, which is more than simply giving marks or grades. Assessment also plays a crucial role in the education process. In Gipps' words, "we must first ask the question '*assessment for what?*' and then design the assessment programme to fit. I take the view that the prime purpose of assessment is professional: that is assessment to support the teaching/ learning process" (Gipps, 1994: 3).

Over time, researchers in education have recognised the power of assessment over teaching in influencing students' learning experiences (Miller & Parlett, 1974; Snyder, 1971; in Gibbs & Simpson, 2004). Through assessment during their learning processes, students may be able to perceive which aspects of the course are most valued. Much of the work they undertake and their approach to learning will be influenced or even determined by such perceptions. The power to influence learning will be even higher when the stake of the assessment is high, such as a professional examination or the final examination for an academic degree. Rowntree even stated that "if we wish to discover the truth about an educational system, we must first look to its assessment procedures" (Rowntree, 1987: 1). Of course, assessment alone may not be able to change every aspect of learning, but it is an effective device to influence teaching and curriculum to

enhance the learning experience. Assessment influences learning by providing students with the motivation to learn, by helping students decide what to learn, how to learn, and learn to judge the effectiveness of their learning (Broadfoot, 1996; Crooks, 1988; in Stobart & Gipps, 1997: 13-16). As Sawyer pointed out:

> High quality education is based upon sound assessment. In effective instructional programmes, assessment provides convincing evidence to the participants that the curriculum goals and objectives are being met. […] Assessment and testing therefore have a pervasive role in educational enterprises. […] assessment provides invaluable feedback on learning and instruction for an entire program of study and serves as a basis for its evaluation – without valid and reliable assessment, the success of a program cannot be gauged accurately (2004: 5-7).

In interpreter education, therefore, it is important that assessment be regarded as an integral part of the training process. The principle is to design and adopt an assessment method that can be aligned best with the overall aims of the training programme, can effectively assess the learning objectives of study programme, and support the development of students' professional competencies, i.e. the interpreting skills and the relevant knowledge about the profession.

Assessment based on the professional knowledge of conference interpreters can inform and enhance curriculum development. There are problems, however, when the method of the interpreting assessment is put under scrutiny by using the fundamental concepts of assessment like *validity* and *reliability* in more established disciplines, such as language testing and educational assessment. This research study attempts to explore and find some answers to those problems.

1.1 Challenges and problems in the interpreting assessment

1.1.1 Seeking a theoretical foundation

In order to systematically investigate the issues of interpreting assessment, some researchers in the field of translating and interpreting have advocated making use of the knowledge of well-established disciplines, such as language testing and educational assessment in general, and seeking insights from them (Sawyer, 2004: 93; Hatim and Mason, 1997: 165-166). Campbell and Hall reviewed the published literatures on the practice of assessing translators and interpreters, and concluded that the field of assessment for translating and interpreting was still "in its infancy", and that the "new kid on the block" could benefit from the solid source of knowledge available in the field of measurement and evaluation (Campbell & Hale, 2003: 221). Sawyer also found that the potential for developing the interpreter assessment was "vast" (2004: 31), and made a substantial contribution on the subject based on the concepts and principles from the fields of language testing and educational assessment (Pöchhacker, 2004: 187).

The concepts of validity and reliability in assessment, i.e. the usefulness and consistency of the tests, have been emphasised throughout research literatures on assessment and test development in all kinds of endeavours. They are regarded as the touchstone of educational and all forms of assessment. The challenges and problems in the current practice of the interpreting assessment mainly are concerns about the test validity and reliability issues, which are in serious need of clarifications (Campbell & Hale, 2003; Hatim & Mason, 1997; Liu, Chang, & Wu, 2008; Sawyer, 2004). These concerns will be reviewed in Chapter 2.

1.1.2 Challenges when assessing an interpreting performance

Unlike translation or other types of tests that may leave a paper trail for assessment purposes, an interpreter's work is ephemeral in nature. After a conference or meeting, on most occasions, there is little record of what the interpreters have done except for the "impressions" that people may have after listening to the interpretation. This makes quality assurance work as well as the interpreting assessment difficult to carry out. In the words of Riccardi, "Interpreting is something evanescent, which vanishes as soon as it is performed. What remains are the impressions received by the audience" (Riccardi, 2002: 116). The impressions that the audience receive in a conference may also be similar to all that the examiners have when assessing interpreters' performances at an examination panel. Although audio or video recordings of the examinees' interpreting performances may be used for examination purposes, the examiners still need to listen to them and judge the performances as if in a live panel examination. Therefore, it is important to understand how this practice of judging-by-impression may affect the validity and reliability of the interpreting examinations.

In fact, assessing a performance of simultaneous interpreting is more complicated than it seems. On the surface, the task of interpreting itself appears simple in terms of its format: listening to a speech (recorded or live) and interpret it into another language. However, simultaneous interpreting is considered by researchers to be one of the most complex human language activities – "a rich and complicated phenomenon in both its cognitive and communication aspects" (Liu, 2001: 86). Measuring a skill performance with a high complexity such as simultaneous interpreting is certainly a challenging job. It is demanding even when the interpreting task is conducted under a controlled examination situation without having to deal with all the external working conditions, such as in a real conference setting.

For example, at a live examination panel, the examiners perceive and judge many components in a simultaneous interpreting performance, such as "the fidelity of the target-language speech, the quality of the interpreter's linguistic output, the quality of his or her voice, the prosodic characteristics of his or her delivery, the quality of his or her terminological usage" (Gile, 1995b: 151). All these were often assessed in real time, such as in the step-by-step description below:

> For a speech consisting of succeeding segments 1, 2, 3..., such an on-line operation would involve listening to segment 1, keeping it in memory while listening to segment 2 and to the target-language rendition of segment 1, comparing the source-language and target-language versions of segment 1 while storing the source-language version of segment 2 and listening to segment 3, etc. (ibid: 152).

Gile suggested that "the short-term memory load in such a sequence is beyond the maximum capacity of most if not all assessors," and believed that "it is practically impossible to monitor all of the original speech and all of its interpretation on site" (ibid). An assessment process such as this involves many qualitative and quantitative decisions that are to be made in real time, and thus the complexity of the examiners' mental activity may be no less than that of the interpreters.

Given the complexity of the task itself when performing simultaneous interpreting (SI) and of the high cognitive demand on the examiners, the judgement as a result is usually made in a holistic and subjective manner, which has raised concerns in the consistency of the judgement process of the examiners[1] (see 1.2.2). Understanding how the examiners exercise their judgement for a better test design can help balancing out the ephemeral nature of simultaneous interpreting in assessment.

[1] Although interpreting output can in principle be recorded and listened to later, thus reducing the cognitive load on assessors, technical and time constraints usually make this difficult to operationalise in a panel examination setting, in which live marking is the preferred method.

1.2 A research gap – rationale and purpose of the study

1.2.1 Scarcity of studies in interpreting assessment

Sawyer (2004) gave the most comprehensive discussion so far on the issues of interpreter education with a dedicated literature review chapter on the foundations of interpreter performance assessment, emphasising the importance of the role of assessment to interpreter education and professionalism. The discussions in his book were set on the wider picture of curricular design, and the results of his case studies showed that "gathering evidence of the validity of interpreting tests is a process of documentation that provides insight on how programs can be optimized and streamlined" (Sawyer, 2004: 231).

Nevertheless, documentation of the actual interpreting examination procedures, especially those for simultaneous interpreting, is difficult to find in the public domain. For example, a search online for the two national examinations of translation and interpreting in China and Taiwan[2], only reveals a skeleton document of "examination outline" that says little about the examinations themselves. The Institute of Linguists in the United Kingdom provides assessment criteria and mark sheets in their Diploma of Public Service Interpreting (DPSI) Handbook (IoL, 1994: 9), but the criteria descriptions are very general in terms of operational practicality (see 2.3.2.c) and are not for conference interpreting. The National Accreditation Authority for Translators

[2] The Examination Outline for Level 2 simultaneous interpreting of China Accreditation Test for Translators and Interpreters (CATTI) is available online at http://bbs.catti.china.com.cn/down/syllabus_EN_SI2.pdf. Accessed 17 May 2009.

Taiwan's Chinese and English Translation and Interpretation Competency Examination (CETICE) Exam Outlines is available at http://www.edu.tw/BICER/content.aspx?site_content_sn=20212. Accessed 17 May 2009.

and Interpreters Ltd in Australia also explains how its accreditation tests work, but recognises that "it is not possible to simulate perfectly in a test the conditions under which an interpreter would normally work" (NATTI, 2010: 8), and only uses one sentence to describe the assessment criteria of consecutive interpreting, which says that "the candidate is required to interpret into the other language almost immediately, providing a structured and accurate rendering of the original" (ibid: 7). Like the DPSI Handbook, this does not explain much about the assessment criteria of the consecutive interpreting test. The American Translators Association (ATA)[3], the main professional accreditation body for translators in the United States, only offers written translation examinations to certify translators, but not interpreters. Two main international organisations for professional interpreters, the American Association of Language Specialists (TAALS) based in Washington D.C., and the Association Internationale des Inteprètes de Conférence (AIIC) based in Geneva, only specify with details the working conditions for professional interpreters and their membership requirements. No official publications of assessment criteria for interpreting examinations, if any, are readily available in the public domain of the professional organisations[4] (also see 2.1.3).

Campbell and Hale carried out a literature review study on the assessment of translators and interpreters and found that "little has been written on interpreting assessment in general, even less is found on any type of assessment as part of training courses" (2003: 216). Pöchhacker also pointed out that within the pedagogical context, few of the assessment issues of interpreter education have been thoroughly treated in the literature (2004: 187). The scarcity of literatures on the topic of interpreting assessment within the educational context indicates a clear gap for research in this area.

[3] ATA web site can be found at http://www.atanet.org/certification/aboutexams_overview.php. Accessed 23 January 2011.

[4] TAALS's standards of professional practice for conference interpreters and translators can be found at http://www.taals.net/standards.php. Accessed 23 January 2011.

AIIC Professional Standards can be found at http://www.aiic.net/ViewPage.cfm/article122.htm. Accessed 23 January 2011

1.2.2 Concerns about subjective judgement and test design

Serious concerns have been raised about how consistent professionals in the field of interpreting can exercise their judgement when it comes to assessing interpreting performances (Sawyer, 2004: 188). Performance assessment has long been criticised as unreliable and in need of systematic study (Campbell and Hale, 2003: 212) and the concerns about the problematic role of professional judgment are mainly due to its subjective nature (Messick, 1989: 91; in Sawyer, 2004: 100). Therefore, proper test instruments and examination procedures are required to facilitate a sound and reliable judgement, and to report the test results by combining the examiners' qualitative and quantitative decisions (Pollitt & Murray, 1996: 74).

However, being a "new kid on the block," research in the field of the interpreting assessment is still at the initial stage of exploration. Many important concepts and essential assessment instruments, such as test construct and assessment criteria, are still underdeveloped. A recent government-funded survey study on eleven interpreter educational institutions in Taiwan, the U.K. and the U.S., found that the current practice of interpreting examinations in the higher education still heavily rely on the professional experience of the teaching staff, and the test design is often subjective and intuitive with little support from empirical studies (Liu et al., 2008).

The emphasis on basing the language test design on empirical evidence gives researchers and test designers crucial indicators toward the enhancement of test design for interpreting assessment. In his case study on an interpreter education institution, Sawyer observed that the interpreter examiners' expertise did not necessarily translate into a high degree of agreement in the exercise of professional judgement, and that fluctuation in judgment is evident (2004: 188). Although Sawyer indicated that his case study cannot be generalised to other examination contexts, there was a lack of

conformity in professional judgement. An urgent call was made for more systematic studies of the assessment procedures of interpreting examinations in order to increase the degree of standardisation for interpreting assessments (ibid: 187-189).

The similarities between language speaking tests and interpreting tests are high in terms of the element of subjective judgement and the requirement of spoken language authenticity in the test input and response; both are performance-based assessment. The challenges faced by test designers of the interpreting assessment may be alleviated by learning from the lessons in the field of language testing. Based on the experiences in language testing, therefore, a strategy of prioritising the problems at hand may be formulated to study the issues of the interpreting assessment. The considerations below are rationales of such a strategy that the present study follows in an attempt to address the concerns and problems in the interpreting assessment.

1.2.3 A starting point to address the problems

A test cannot be valid if it is unreliable, i.e. inconsistent test results are useless. Any well thought out examination criteria, procedures and test instruments will be of little value in test reliability, and therefore validity (see 2.3), if the examiners do not use them in a consistent manner or if the design of the instrument itself makes it difficult to be used consistently. The language testing community realised this and conducted empirical studies to systematically improve the assessment tools (such as rating scales) and procedures, especially for those that require more of the examiners' subjective judgement as in writing and speaking tests.

Many reliability studies of language testing also identified the examiners themselves as a source of measurement error (Alderson, Clapham, & Wall, 1995; Bachman, Lynch, & Mason, 1995; Fulcher, 2003; Lumley & McNamara, 1993; Luoma,

2004). These errors in measurement have a subtle way to influence the results of many performance-based assessments, making the assessment procedure become unreliable and threatening the validity of the test (Eckes, 2005: 197). Test instruments, such as rating scales with specific assessment criteria, and training of examiners are often used to help reduce the subjectivity in the assessment and make the examiners more consistent in marking a performance.

In language speaking tests, however, Fulcher pointed out that many of the descriptors in the rating scales were thought up "for the sake of creating consistent looking scales that have little empirical basis", and suggested that the development of rating scales in fluency and accuracy should be based on a recorded corpus of student talk (Fulcher, 1993; in Pollitt & Murray, 1996: 76). Pollitt and Murray further argued that "descriptors should closely match what the raters perceive in the performances they have to grade", and that "the starting point for scale development should surely therefore be a study of the perceptions of proficiency by raters in the act of judging proficiency" (1996: 76). These experiences in language testing provide valuable lessons for the study of test design for the interpreting assessment.

1.3 An exploratory research journey – aim and objectives

As argued in 1.2.3, even well-thought out test design, criteria, test instruments etc. will have little value if they are not used consistently. To achieve a higher degree of test reliability and to develop more practical test instruments such as rating scales and performance descriptors for interpreting examinations, it is reasonable that the starting point should be a study of the perceptions of examinees' proficiency by examiners in the act of judging proficiency. Also, for those interpreting examinations that are judged by a panel of examiners, the interactions among examiners in a panel can only be understood better when we know more about how individual examiners form their opinions before entering the discussions.

Therefore, taking the background and rationale outlined above, and as a starting point, the overall aim of this study is

> to explore and understand how examiners perceive the interpreting performances and make judgments in a simultaneous interpreting examination.

In an ideal examination situation, the judgements of different examiners should be consistent. Due to subjective judgements and potential test errors, nevertheless, this study hypothesizes that it is more likely to have inconsistencies in an interpreting examination, such as the findings in Sawyer's (2004) case study (see 1.2.2). Therefore, in achieving the above overall aim, this study sets itself the first objective as below:

1. To identify and determine if there are different judgement patterns of the examiners when assessing simultaneous interpreting

In their study on language testing, Pollitt and Murray 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 (1996:88). By contrast, Sawyer's case study found that there were clear variations in interpreter examiners' professional judgements (2004: 188), and that the interpreter examiners "make numerous references to various criteria for assessment," which were "fuzzy" (ibid: 185). These observations of the examiners' judgements further prompt the following objectives in achieving the overall aim of this research study:

2. To discover whether interpreter examiners and non-interpreter examiners can achieve similarly consistent judgements in a simultaneous interpreting examination
3. To elicit and understand what the examiners' most important assessment criteria are when judging students' interpreting performances
4. To ascertain the relations between the examiners' application of assessment criteria and their judgement results of student interpreters' performances

and finally in relation to the above research questions,

5. To understand the examiners' assessment behaviours that may lie behind their consistent and inconsistent judgements

In achieving the above aim and objectives, on the theoretical front, it is hoped that this study can contribute to a better understanding of how examiners exercise their judgement in the simultaneous interpreting examinations, including the assessment criteria they used to measure the construct of interpreting, which can lead to the formulation of a conceptual model as a foundation for explaining how the interpreting examinations work. On the practical front, it is hoped that with a better knowledge of the workings of the interpreting examinations, i.e. the conceptual model, better interpreting examinations may be developed in the future with more practical and reliable instruments and procedures for assessment.

1.4 An outline of the research study

After the introductory chapter, this thesis follows the outline below to report the research findings on the workings of the interpreting examinations.

Chapter 2 aims to give the theoretical framework of present study by reviewing literatures in two main areas: (1) the professional quality assurance in the field of conference interpreting and its relation to the interpreting assessment in the context of education, and (2) the essential concepts and knowledge in the discipline of language testing, mainly the concepts of test validity and reliability. Theoretical considerations of research methodology are also presented at the end of Chapter 2.

Chapter 3 explains the study methodology. This research study adopted a multi-strategy approach in collecting research data (Bryman, 2004: 452). Quantitative as well as qualitative methods were employed to gather and analyse data. Thurstone's Method of Paired Comparison was used to monitor the consistency level of examiners, and during the comparison process the examiners were asked to give verbal comment and interviewed to obtain qualitative data for analysis.

Chapter 4 presents and analyses the quantitative data from the paired comparison method, and formulate a framework for analysing the qualitative data.

Chapters 5 and 6 present and analyse the qualitative data: the use of assessment criteria and the examiner behaviours.

Chapter 7 discusses and clarifies the intricate relations of the test constructs of interpreting based on the study findings, and formulate a conceptual model of the interpreting examinations.

Chapter 8 concludes the study and makes some suggestions for future studies.

CHAPTER 2

A New Kid on the Block

2.0 Overview

Sawyer pointed out that "Interpreting Studies […] is beginning to discover constructs from the field of assessment, measurement theory and language testing. At this stage, little literature on validation is extant" (Sawyer, 2004: 34). Hence, a general framework of assessment for interpreter education needs to be developed by using the knowledge from more established disciplines such as language testing and educational assessment (Campbell & Hale, 2003: 221).

Professional judgement and practice in the field is an important source for developing and validating performance tests such as interpreting assessments. Test developers need to draw knowledge and experience from the profession of interpreting for test constructs in order to make the tests valid. Therefore, the state of interpreting quality assurance in the profession will first be reviewed (2.1) to gain insights into how professionals as well as the interpreting service users perceive the quality of interpreters' work, and how the knowledge in the field of the profession may be applied to assessment in the educational context (2.2). Then, important concepts in testing theory including the theoretical foundation of the research study, will be reviewed, which will focus on issues surrounding test validity and reliability, and their implications to interpreting assessment (2.3). Some test design methods, such as test specifications that may help develop more useful and dependable examinations are reviewed in the latter part of the chapter (2.4). The final section reviews the rater effect and theoretical considerations of the research methodology for this study (2.5).

2.1 The nature of quality assurance for conference interpreting

Conference interpreting as a profession is relatively young. Research on quality assurance of interpreting services dates back only to the 1980s (Pöchhacker, 2004: 153), and came to the fore in interpreting studies in the 1990s (Pöchhacker & Shlesinger, 2002: 296). At first, quality assurance was based on the assumption that the professionals, i.e. interpreters and their peers would know what was good (ibid). However, due to the complex nature of conference interpreters' work, in the early days many of the professionals and researchers in this discipline found it difficult to describe or explain quality in interpreting systemically with clarity, and regarded it as an "elusive concept" (Shlesinger, 1997: 123). Interpreters often perceived and characterised the quality of their work in practice. Their reflections were then reported and published by using "intuitive" and "sweeping statements" (ibid: 123-124), and their research approaches were regarded as "experiential and impressionistic" (Sawyer, 2004: 20).

Nevertheless, over time, studies in the field of interpreting started to use different ways of looking at quality in interpreting, including not only the perspectives of the interpreters, but also those of the listeners, i.e. the 'users' of the interpreting service (Pöchhacker, 2001). The quality issues of interpreting were explored through more systematic approaches by using verifiable data for research, and theoretical models for the quality standards of interpreting were formulated. In the sections below, a snap shot of these studies and their implications are reviewed as the foundation to discussing the validity and reliability issues concerning assessing simultaneous interpreting in this research study.

2.1.1 User expectation survey studies

In business marketing, the concept of quality is measured by comparing the actual service delivered (in this case, interpretation) with the expectations of the customers (in this case, the audience of the interpreting service). Conference interpreting being a service industry by nature, it is natural that the pursuit of service quality should begin with the understanding of the customers' needs and expectations. The important role of the audience in interpreters' work has been emphasised by interpreters since the emergence of the profession. Depending on the characteristics of various work situations, such as political, scholastic, literary, and the size of the audience, conference interpreters noticed that they need to adjust the way they use the target language to facilitate the communication, and that the audience's point of view should play an important role in judging their interpretations (Herbert 1952, Gold 1973, Seleskovitch 1986 in Kurz, 2001: 395).

Kurz regarded this view as the user-oriented professional standard and used it as a measure when evaluating the quality of interpreting, emphasising the importance of understanding the audience's expectations. This research theme helped build up the user expectation profiles of interpreting services and became a fruitful line of research (Kurz, 2001; Pöchhacker, 2001: 415).

As one of the earliest empirical studies on quality issues of conference interpreting, Bühler (1986) conducted a questionnaire survey to identify conference interpreters' expectations of the quality of their interpretation output in relation to the needs of those who use their services. This study was based on the assumption that experts' expectations would also be consistent with those of the users (Pöchhacker & Shlesinger, 2002: 295). The assumption and the survey findings gave an indicator to the understanding of how the interpreters themselves would expect to see the quality of

their own work, and suggested some quality criteria for the interpreters' output, such as accent, voice, fluency, logical cohesion, sense consistency, completeness, grammar, terminology, and style. (Bühler 1986 in Kurz, 2001; Pöchhacker, 2001, 2004: 153).

Inspired by Bühler's survey study, Kurz carried out a series of survey studies on end-users of different backgrounds in various conferences (1989, 1993, 1994, 1996 in Kurz, 2001). Kurz adopted the criteria in Bühler's survey questions in her studies, and found that in terms of importance, the average ratings of criteria by end-users were consistently lower than those of the conference interpreters; only three criteria – sense consistency, logical cohesion and correct terminology – had similar ratings to Bühler's results (Kurz, 2001). This showed that the conference interpreters have a higher expectation of what they do than the end-users have, and the high standard of conference interpreters in what they do contributed to shaping and defining the emerging profession. Kurz's studies also supplied an important finding that user's expectation profiles were different from one another according to their professional background, i.e. people go to different conferences with different expectations. This finding prompted more studies to look into the quality issues of interpreting from a multi-dimensional point of view.

Moser conducted a large-scale survey study between 1993 and 1994, utilising 94 interpreters to interview over 200 people who had participated as a speaker or as an audience in 84 international conferences around the world. The survey findings generally confirmed the importance given by users to criteria such as completeness, clarity of expression, and terminological precision (Moser, 1995), but it identified other factors that affect the user expectation of conference interpreting service.

For example, it was found that user expectations tended to vary considerably depending on meeting type. The larger the conference, the more information the audience needed to take in, and therefore, the greater desire for concentration on the

essentials of messages. This user requirement for succinctness usually has to be fulfilled by conference interpreters when the audience need to rely on simultaneous interpretation to receive information. Conference topics also have some influence on the users' requirement of interpretation: the more technical the conference, the greater the users' desire for completeness and literalness, such as in the use of terminology. In addition, factors such as age, gender and previous experience of using simultaneous interpreting also may affect the users' expectations of the interpreting service (Moser, 1995).

The findings from these user survey studies indicate that user expectation of the quality of interpreting service changes, depending on user background and circumstances of the conferences. Therefore, investigating the quality of an interpreting performance appears to demand an approach that is multi-situational and multi-perspective.

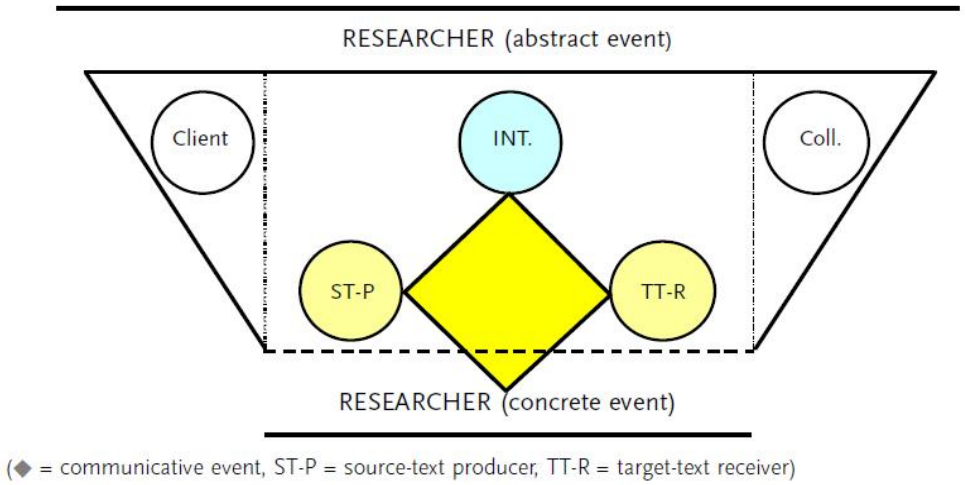
2.1.2 Multi-dimensional perspectives on quality in interpreting

In spite of the findings, the value of these user surveys has also been questioned. The user surveys are said to be insufficient to determine interpreting quality because they show only one limited aspect – the listeners who only understand the target language – in an interpreter-mediated communication situation (Kalina, 2005: 776). Furthermore, the survey studies seemed to be concerned more with the *ideal* quality in interpreting, and the users' hypothetical preferences might not reflect what actually happened when they had listened to a real interpreter. Therefore, a shift of attention to quality under the circumstances of observation was called for by some researchers into the study of quality in interpreting (Bühler 1986, Moser-Mercer 1996 in Pöchhacker, 2004: 154-156).

Here we shall give a snapshot of research studies that responded to this shift of attention. Most of the studies remained survey-based, but they were conducted with different focuses, such as the linguistic criteria of the interpreters' output, the aspects of the role of the interpreters, meeting types as well as the expectations from a source speaker's point of view (Kopczyński 1994, Marrone 1993, Vuorikoski 1993 in Kurz, 2001; Pöchhacker, 2001). Kopczyński identified a dual perspective in looking at the quality of a conference interpreter's work, which he called *linguistic* (the output interpretation quality) and *pragmatic* (quality as seen in a context-dependent perspective) (Kopczyński, 1994: 87). Pöchhacker discussed the interdependencies between the different actors and texts in an interpreted communication event (Pöchhacker, 1994). Kalina (1995) suggested considering the interests and motivations of different parties to a communication act (1995 in Kalina, 2005: 775). Shlesinger argued for the need to scrutinise the output quality on three different levels, namely the intertextual level (sense consistency between source and target texts), the intratextual level (quality of the target language output) and the instrumental level (the usefulness of the interpretation) (Shlesinger, 1997).

The multi-perspective approach in determining the quality of interpreting attracted much interest, including the above research studies, and has been widely discussed in the literature on quality in interpreting (Sawyer, 2004: 117). This led to theoretical modelling that can explain better the complex notions of interpreting quality and be used as a framework that further research studies may be based on. For example, Pöchhacker (2001) used a figure (reproduced here as Figure 2-1) to illustrate the multi-perspective themes of the quality in interpreting. In Figure 2-1, the interpreter (INT.), the speaker (ST-P, i.e. source-text producer) and the listener (TT-R, i.e. target-text receiver) make up the core communicative event with the Client and Colleague (Coll.), i.e. fellow interpreters in the booths or in the audience, on the sides as additional

Figure 2-1 Perspectives on quality in interpreting



Source: (Pöchhacker, 2001: 412)

roles who may also take a position in assessing the quality of interpreting. In addition, the two researcher positions in the figure show the analytical distinctions underlying the study of quality in interpreting. The researcher at the top is an outsider of the communicative event, who takes a more abstract overview of an interpreting event; whereas the one at the bottom has more direct assess to, or is a part of, the concrete communicative event. In the latter case, the recordable product (i.e. the interpretation) and the overall process of communicative interaction may be directly observed (often in real time) and analysed by the researcher, and so empirical evidence can be produced. These product- and process-oriented approaches may, therefore, facilitate a more systematic understanding of the quality-related issues in interpreting, such as standards and criteria for assessment (Pöchhacker, 2001: 412), and how the assessment work is carried out, such as the current study.

Thus, Pöchhacker's multi-perspective figure is also useful to illustrate the current research study on the testing issues concerning simultaneous interpreting examinations. The researcher of this study takes the position at the top in Figure 2-1 to observe and

analyse the interactions (the *abstract* event) between three elements in a simultaneous interpreting examination (the *concrete* event), namely the source speech, the interpretation output and the examiners; whereas the examiners take the position of the "researcher" at the bottom in Figure 2-1, who participated in the concrete event because they also observe the interactions between the source speech and the interpretation output when judging the interpreting performance. To certain extent, the examiners also play the role as the audience because they listen to the interpretation at the same time. The design of the present research study is explained in details in Chapter 3.

2.1.3 Professional standards and interpreting quality assurance

Given the multi-perspective views, expectations and approaches mentioned above, how can the quality standard for simultaneous interpreting be described, or for that matter, be measured? Association Internationale des Interprètes de Conférence (AIIC), the professional organisation for conference interpreters founded in 1953, has been successful in setting uniform standards for simultaneous interpreting at international conferences throughout the world (see Keiser 1999 in Pöchhacker, 2004). In order to ensure the best possible quality for interpretation, the AIIC specified a range of work conditions for conference interpreters in its Code of Professional Ethics (AIIC, 2009). Detailed guidelines of professional practices regarding the contractual support to conference interpreters were also drawn up in the AIIC Professional Standards, the purpose of which is "to ensure an optimum quality of work performed with due consideration being given to the physical and mental constraints inherent in the exercise of the profession" (AIIC, 2000).

Nevertheless, in spite of giving detailed requirements of the working condition for its members, the AIIC has been criticised as being "silent about issues of role and

performance quality" of conference interpreters in its Codes (Pöchhacker, 2004: 164). The AIIC does include a Code of Honour that imposes an article of non-disclosure of information on its members by "the strictest secrecy" (AIIC, 2009). Other than that, however, the interpreting community has been said to be still reluctant to challenge its own codes and regulations (Angelelli, 2006: 175), such as giving explicit statements to specify the quality of interpreting work that its members should provide. One possible reason why the interpreting community seems reluctant to do so may be revealed in Kalina's comment below:

> Interpreters and trainers feel that they can assess the quality of colleagues or trainees intuitively, on the basis of their experience and professionalism, but they are unable to express their subjective judgments by objectively measurable standards. (2005: 768).

In other words, the concept of quality in interpreting remains largely vague and subjective among the interpreting community.

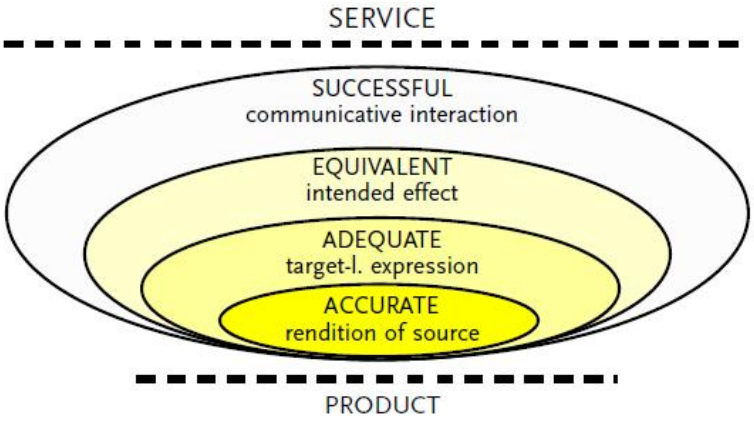
In addition to the tricky problem of dealing with subjective judgement (see 2.3.3.a), the challenge of expressing "objectively measurable standards" of interpreters' work is significantly complicated by the multi-dimensional perspectives that were discussed above, and by the requirements of the interpreters' work conditions, such as the arrangement of the interpreting facilities, the distribution of conference documents, and the team formation of conference interpreters (AIIC, 2009).

The sheer number of variables such as the above may be the reason why the quality of conference interpreting seems "elusive" as Shlesinger put it (1997: 123), and difficult to analyse. A generally accepted quality model is yet to be found for conference interpreting, or any type of interpreting (Kalina, 2005: 768).

2.1.4 Pöchhacker's model of quality standards for interpreting

In order to capture and clarify the elusive concept of interpreting quality, Pöchhacker presented a model of the quality standards for interpreting (reproduced here as Figure 2-2) which shows a "fundamental duality of interpreting as a service to enable communication and as a text-production activity" by the interpreters (2001: 412).

Figure 2-2 Quality standards for the product and service of interpreting



Source: (Pöchhacker, 2001: 412)

Pöchhacker's service–product duality model echoes Kopczyński's (1994) linguistic–pragmatic dual perspective; interpreting essentially is a linguistic service and its product, i.e. interpretation, needs to be pragmatic. Pöchhacker's model further makes finer distinctions between four elements of quality in interpreting. Expanding outward like multiple ripples, Pöchhacker's model lays out the multi-dimensional criteria or standards for interpreting with *accurate rendition of source* at the core, encompassed by three outer layers of *adequate target language expression, equivalent intended effect*, and *successful communicative interaction* (see Figure 2-2).

Using this quality model, performance expectations from different perspectives such as those from the user survey studies can be mapped onto the different layers. For example, the three highly-correlated criteria in Bühler (1986) and Kurz (2001) – sense

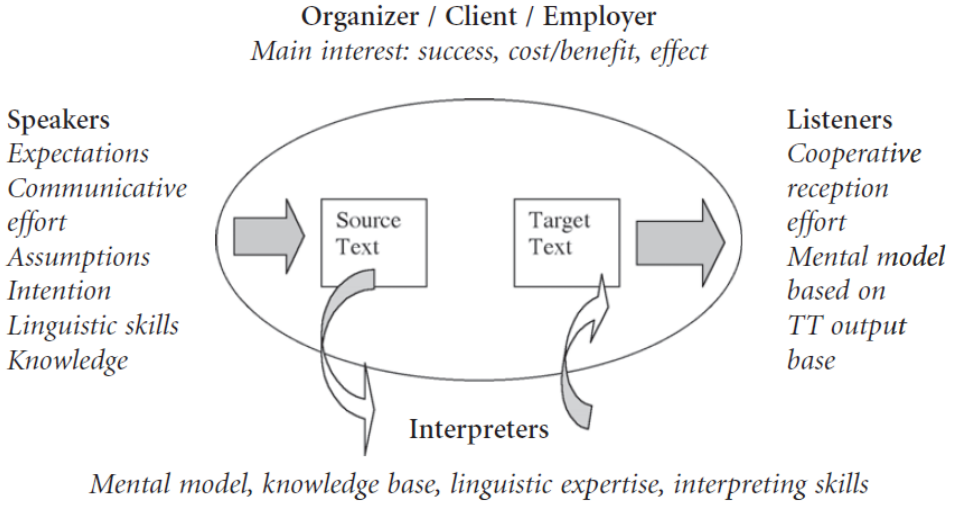
consistency, logical cohesion, correct terminology – are at the core in the accurate rendition of the source, whereas the other criteria, such as accent, voice, fluency and style, etc., fall in the next two layers of adequate expression and equivalent intended effect. And finally, the outermost layer that depicts a successful communicative interaction, say in a conference, includes all the other situational variables, such as the types and subjects of conferences as well as the participants' backgrounds (see Moser, 1995). The overall working conditions that the AIIC recommended can thus also be placed in the outermost layer of Pöchhacker's (2001) quality model, such as effective support from the conference organisers, the quality and arrangement of the interpreting facilities, the distribution of conference documents, and the team formation of conference interpreters (AIIC, 2009). If those working conditions are adequately provided, they will help the interpreters to achieve an optimum quality of work in a conference, i.e. enhancing the quality of the inner layers and achieving an optimum quality of the communicative interaction in Pöchhacker's model.

2.1.5 Kalina's integrated approach to quality assurance for interpreting

Just as shown in Pöchhacker's standards model above, interpreting quality includes aspects that cannot be explained by simply comparing the source and target languages. The multiple perspectives of different parties to a communication act need to be taken into account, such as their interests and motivations as well as the purpose of the communication, when judging the quality of interpreter output.

From a psycholinguistic perspective, interpreted communication has been regarded as a complex discourse-based mental process (Kohn & Kalina, 1996; in Kalina, 2005). The illustration of the process is reproduced here as Figure 2-3. It shows clearly that an interpreter-mediated communication involves production and comprehension by the

Figure 2-3 Bilingual, interpreter-mediated conference communication



Source: (Kalina, 2005: 776)

different participants in communication. With the source and target texts at the centre, the multiple participants in the communication act, i.e. Speakers, Listeners, and Organizer/Client/Employer, all have different interests and expectations of the communication act, which is facilitated by the expertise of the interpreters.

As seen in Figure 2-3, therefore, interpreting quality assurance (QA) should consider not only the output and effect of the interpreter's rendition of the source text, but also the structures, processes and conditions of the communication event that are likely to have a bearing on interpreting quality. All these considerations would help describe the strategic choices available to interpreters when undertaking an interpreting assignment. Thus, an effective interpreting QA scheme needs to cover what happens before and after the interpreting act itself since these phases may have a significant impact on interpreter output (Kalina, 2005: 776-778). It should capture and record the strategic choices that are available to the interpreters and how they make the choices, so that the interpreting service can be evaluated for what they are, and good as well as poor interpreting quality can be identified as such (Moser-Mercer 1996, Kalina 2002, Mack 2002 in Kalina, 2005).

Considering the above, Kalina proposed an integrated framework for interpreting quality assurance. The QA framework consists of four phases as in an interpreting assignment: pre-process (arranging the contract and preparing for the assignment), peri-process (preparation just before the assignment), in-process (interpreting at the conference), and post-process (debriefing). The QA framework is in the form of a datasheet to record 77 items that requires the attention of the interpreters (Kalina, 2005: 779-781), and is a macro-level approach, which covered almost all aspects of a typical conference interpreting assignment. The datasheet was designed mainly from the interpreter's point of view, but it also provides useful information and reminders for clients, speakers and conference organisers so that all those involved understand which limitations are intrinsic to a mediated multilingual communication and what they can do to achieve an optimum result (Kalina, 2005: 782).

Apart from professional QA endeavours as reviewed so far, another important task is to determine how the performances of the interpreters and interpreter students can be assessed and measured in a valid and reliable manner by employer recruiters and interpreter training programmes. Where concepts such as the quality of interpreting have yet to be illuminated in the field, within the educational institutions it has to be researched and explained systematically for the sake of pedagogical development, of the reliability of the assessment, and the credibility of the institutions that carry out the interpreting examinations. Informed by the professional experience, it is therefore also worth noting that the four requirements listed under the Interpreters in Figure 2-3 – *mental model, knowledge base, linguistic expertise, interpreting skills* – may be useful to defining the test constructs of interpreting.

In the following sections, concepts in testing (such as validity, reliability, construct) and specific issues that surround the assessment of trainee interpreters' performances are reviewed as the theoretical foundation for the current research study.

2.2 From professional quality assurance to educational assessment

Before looking into the issues that surround the assessment of interpreter performance in training, some clarification needs to be made of differences between quality assurance in the profession and quality assessment in the context of education. Merriam-Webster (2009) defines *quality assurance* as

> a program for the systematic monitoring and evaluation of the various aspects of a project, service, or facility to ensure that standards of quality are being met.[5]

Therefore, quality assurance for interpreting can be said to be a set of planned activities or processes that ensure in a systematic and reliable fashion that the product (interpreting services) satisfies the customers' expectations (audience, speakers, employers, conference organisers, etc.). The integrated framework that Kalina proposed (2.1.5) is an example of such planned activities for interpreting quality assurance.

In contrast, the work of quality assessment in the context of education may be viewed in two aspects: the quality of teaching and the quality of students' learning outcome. The quality assessment in education usually takes the form of various types of tests and examinations. It is best to describe assessment in education as an action, usually in measurable terms, to determine the importance, size, or value of a skill or subject knowledge. Gipps describes assessment in education as

> a wide range of methods for evaluating pupil performance and attainment including formal testing and examinations, practical and oral assessment, classroom based assessment carried out by teachers and portfolios (Gipps, 1994: vii).

[5] quality assurance. (2009). In Merriam-Webster Online Dictionary. Retrieved May 30, 2009, from http://www.merriam-webster.com/dictionary/quality assurance

Using some assessment criteria, teachers measure students' performances in the examinations to ascertain their learning achievements, the results of which in turn reflect certain aspects of the teaching quality.

Assessment needs to be based on a process of validation, which requires gathering evidence to see whether it can support what an assessment aims to assess (i.e. the quality of learning outcome) and what the training programme claims to achieve (i.e. the quality of teaching). And the source where the evidence can be obtained for such validation is usually from the professional quality standards in the field. Their relationship is reviewed next in 2.2.1.

2.2.1 Professional standards and educational assessment

In the case of interpreting, professional quality assurance and educational assessment are partially distinct but not discrete. The discussion of quality in the field can inform the design and development of assessment methods in the training programmes (Sawyer, 2004: 99). The former enhances the quality of professional practice, whereas the latter not only measures the interpreter students' educational attainment but also often serves to ensure that those who join the profession are suitably qualified. This suggests a strong interdependent relationship between professional standards and assessment in interpreter training (Pöchhacker, 2004: 187).

The design and development of interpreting examinations can benefit a lot from the QA endeavours for interpreting work. For example, the problems that are central to the assessment of interpreting include sense consistency and delivery adequacy as shown in the two core layers of Pöchhacker's model of quality standards for interpreting (Figure 2-2). From a vantage point with the multifaceted perspective of all the parties in the multilingual communication event, Kalina's macro-level framework (2.1.5) can help

researchers and test developers get insightful views of the micro-level problems in the assessment of interpreting. Crucially, constructs for testing may be identified from these quality models and frameworks to formulate useful assessment criteria for the interpreting examinations, such as the aforementioned ability requirements that were found under the Interpreters in Figure 2-3 – *knowledge base*, *linguistic expertise*, and *interpreting skills*, which can be made operational in testing by understanding more about the *mental model* of the interpreters.

Another example to illustrate the interdependent relationship between professional standards and educational assessment is Riccardi's discussions of the assessment criteria for interpreting. Riccardi (2002) distinguished two types of assessment criteria: *macrocriteria* for professional interpreters and *microcriteria* for student interpreters. The professional macrocriteria include equivalence, accuracy, appropriateness and usability, taking into full account user expectations as well as the restrictions of the communication act. On the other hand, the educational microcriteria consist of fourteen categories, including register, omissions, deviations in content, successful solutions, and so on. The microcriteria were derived from studies of interpreting quality in the professional field and from Riccardi's own experience as an interpreter teacher. Riccardi indicated that the macrocriteria of the professionals ultimately subsume the microcriteria that students are learning (Riccardi, 2002).

From the above, the interdependent relationship between professional practice and the assessment of student interpreters is clear. The relationship hinges on the validity of the foundation that the study and practice of professional interpreting can provide for the training and assessment of interpreters in the educational institutions. To this end, therefore, Riccardi called for further research in order to determine scientifically the validity of her proposed assessment tool (Riccardi, 2002: 125). This brings us to the issues of interpreter performance assessment in the educational context.

2.2.2 Types of assessment: formative vs. summative

In the *Prelude* to this thesis, a typical scenario of an interpreting examination was presented, in which student interpreters' simultaneous interpreting performances were assessed in an educational institution. Some questions and doubts regarding the operation of interpreting examinations have been raised, which are similar to the interpreter teacher's reflections in the *Prelude*. Hatim and Mason were among the first to point out such "unease felt by many at the unsystematic, hit-and-miss methods of performance evaluation" across many translator and interpreter training institutions (Hatim & Mason, 1997: 165-166). Assessment in interpreter training is a highly complex subject (Pöchhacker, 2004: 187) and sometimes a very controversial issue (Riccardi, 2002: 116). It is complex because of the multi-dimensional nature of the interpreting job as well as the interdependent relationships between professional standards and educational considerations; it is controversial because subjective judgement is involved in the assessment process. These are perhaps the reasons why "hit-and-miss methods" are all too often used for the interpreting examinations.

In education, depending on the purposes of the test, two types of assessment were identified as *formative* and *summative* assessment (Scriven, 1967). When considering the methods of assessment to meet the aims and objectives of a translator and interpreter training programme, Hatim and Mason (1997) pointed out that a clear, initial distinction needs to be made between these two types of assessment. The main goal of formative assessment is to support the learning process by using the information obtained from the assessment as feedback to students, helping them to improve "by short-circuiting the randomness and inefficiency of trial-and-error learning" (Sadler, 1989 in Gipps, 1994: 125). In contrast, summative assessment aims to give evidence for decision-making, such as deliberating on students' fitness to proceed to the next level of study, or to be

awarded certification or a professional qualification (Hatim & Mason, 1997: 166). In other words, summative assessment "provides information about how much students have learned and how well a course has worked" (Gipps, 1994: vii; in Pöchhacker, 2004: 106). As discussed earlier, since assessment also gauges the quality of teaching, high fail rates in the final examinations of a course may also be regarded as "an indication that the educational objective of the program are not being met" (Sawyer, 2004: 113).

Generally speaking, summative assessments are more visible at the end of a term or a study unit for decision-making purposes; whereas formative assessments usually happen during the course of teaching, which may be more frequent and interactive in order to identify learning needs and adjust teaching appropriately (CERI, 2005: 21). However, a formative assessment that is not being used for supporting learning is not for truly formative purposes even when it takes place during the course of the study. The key difference is the purpose and effect, not the timing of the assessment (Gipps, 1994: 125). Based on this view point, for interpreter training Sawyer suggested that the *intermediate* assessment of apprentice interpreters could be both formative and summative – "formative in that feedback is given on a student's work, which guides decision-making for continuation in the degree program, and summative in that learners demonstrate baseline competence on specific occasions" (Sawyer, 2004: 113).

For both types of assessment, they usually require the demonstration of a certain level of predictive power of students' future performance, and provide evidence in the demonstration so that the test results can be trusted, i.e. the consideration of test validity and reliability. In this regard, Gipps suggested that

> If assessment is to be used for certification or accountability then it needs an adequate level of reliability for comparability purposes. If however, the assessment is to be used for formative purposes, validity […] is highly important and reliability is less so. (Gipps, 1994: 137)

This section has reviewed the relationship between professional standards and assessment in interpreter training. The pursuit of professional quality assurance in the field can offer valuable input and validation evidence for interpreting training programmes, and in turn the educational institutions serve as a gateway to ensure that those who join the profession are suitably qualified. The important roles and types of assessment for the interpreter training were also reviewed. Depending on the purposes of assessment, different types of assessment method should be considered. There are different considerations when choosing an assessment method, and they are often related to the fundamental issues of test validity and reliability, which will be reviewed next in 2.3 as the theoretical foundation to facilitate further discussions in this study on the issues of the interpreting assessment.

2.3 Theoretical foundation – test validity and reliability

Simply put, validity of a test refers to the meaningfulness of the scores, and defines a broad scope of concerns for their use in a community or a wider social context (Luoma, 2004: 184). If a test is not valid for its purpose of use, the meaning of the test scores is void or even misleading. Validity should be, therefore, of "central concern to all testers" (Alderson et al., 1995: 170).

As for reliability, it refers to the considerations in evaluating a test design, including its procedure and report of test results. In particular, a test should demonstrate consistency in its operation so that it can be considered a trustworthy form of measurement. Whatever the standards may be, they should be the same for all persons being assessed and for all persons asked to make judgments on different occasions and over time (Bailey, 1998; Hamp-Lyons, 1991).

This section first reviews the concept of test validity as understood in the field of educational assessment and language testing, and its implications for the interpreting assessment (2.3.1 & 2.3.2), followed by a discussion of reliability (2.3.3) and its relation to validity in testing (2.3.4). A challenge to the development of interpreting examinations is also discussed at the end of this section (2.3.5).

2.3.1 The concept of test validity

In educational assessment, Gipps summarised the definition of validity as "the extent to which a test measures what it is designed to measure" and added that "if it does not measure what it purports to measure, then its use is misleading" (1994: 58). These statements imply two things in test validity: the former statement is about the

characteristics of a test, for instance a written translation examination or an oral interpreting examination, and the latter is about the *inferences* that can be appropriately drawn from a test, such as the examinees' ability to translate or interpret at certain levels or within a specific context, i.e. the usefulness of the test. The characteristics and inferences of a test are closely related to each other so they are commonly referred to together as test validity (Salvia & Ysseldyke, 1995: 162). In other words, test validity is a concept that links what is tested and the usefulness of the test results.

Some authors also took the view that the social consequences also need to be considered when drawing inferences from testing (Messick, 1989 #155: 19 in Salvia & Ysseldyke, 1995: 162). For example, international airline pilots need to be tested to make sure that they can communicate in English with air traffic controllers in other countries. Therefore, the social consequence of the pilots' English test also relates to the safety of the passengers and the dependability of the airline industry. Another example is IELTS and TOEFL, whose test results every year determine whether or not thousands of candidates can enter a university to study in an English-speaking country. Therefore, it is expected that evidence is provided to support test interpretation and potential consequences of test use (Gipps, 1994: 59), i.e. to establish test validity.

2.3.2 Methods for test validation

Alderson et al. pointed out that "It is not enough to assert, 'This test is valid' unless one can answer the follow-up question: 'How do you know?' and 'For what is it valid?'" (1995: 170). Evidence needed to be gathered to answer these questions in order to establish a test's validity, which should be done in as many ways as possible (ibid: 171). The process of gathering evidence for determining the appropriateness of test inferences, i.e. *test validation*, is an iterative process that "link concepts, evidence,

social and personal consequences, and values" (Cronback, 1988: 4; in Sawyer, 2004: 95). Then, the gathered evidence can be categorised to facilitate explanation of the test validity, which is "an artificial device to explain the possibilities; there are not separate validities" (Salvia & Ysseldyke, 1995: 162). So the *types* of validity are actually labels attached to different evidence that help to explain and support the "*adequacy* and *appropriateness* of *inferences* and *attitudes* based on test scores or other modes of assessment" (Messick 1989: 13; as quoted in Gipps, 1994: 59). In other words, the types of validity "are in reality different 'methods' of assessing validity" (Alderson et al., 1995: 171).

Traditionally in educational assessment there are four types of validity: predictive, concurrent, construct and content validity, which Gipps summarised as follows:

> *Predictive* validity relates to whether the test predicts accurately or well some future performance. […] *Concurrent* validity is concerned about whether the test correlates with, or gives substantially the same results as, another test of the same skill. […] *Construct* validity itself relates to whether the test is an adequate measure of the construct, that is the underlying (explanatory) skill being assessed. […] *Content* validity concerns the coverage of appropriate and necessary content, i.e. does the test cover the skills necessary for good performance, or all the aspects of the subject taught? (1994: 58-59)

Together the concurrent and predictive validity are also referred to as *criterion* validity "because they both relate to predicting performance on some criterion either at the same time or in the future" (ibid: 59). In language testing, Alderson et al. referred to this type of validity as the *external* validity because the evidence was "gleamed from outside the test" and the content validity was in the category of *internal* validity because the evidence lies within the test itself (1995: 171).

In order to avoid confusion in this study, the term "external validity" is adopted because the word "criterion" is used in a different context in this thesis, such as the assessment criterion, or criteria, in the interpreting examinations.

It was found that of all the types of validity, however, evidence about only one or two of the types of validity were usually provided for test development. Therefore, researchers in the field of assessment gradually began to take the view that validity is a unitary concept with *construct* as the unifying theme (Messick, 1989; Cronbach, 1988; in Gipps, 1994: 59). "All types of validity are subsumed under one concept", and validation requires test developers to specify whether the evidence supports construct-, content-, or externally-referenced validity (Sawyer, 2004: 96).

As the discipline of assessment for interpreters is still in its infancy, there is a "paucity of discussion on the central topics of validity and reliability", showing a large knowledge gap in these areas (Campbell & Hale, 2003: 218). In the absence of a comprehensive review of assessment validity issues in the community of interpreter educators, Sawyer offered a literature review discussion on the validity issues in relation to interpreter education (2004: 96-101), which included some sample statements for the validation of interpreting assessment as below:

External *(Criterion)* validity	Graduates are able to work in their market sector Graduates can pass same or similar tests again in the future, including those administered in industry
Content validity*	Consecutive interpretation Liaison interpretation Simultaneous interpretation Simultaneous interpretation with text
Construct validity	Ability to - Interpret with faithfulness to the meaning and intent of the original - Use appropriate language and expression - Apply world knowledge and knowledge of subject matter - Demonstrate acceptable platform skills and resilience to stress

(Sawyer, 2004: 97)

**Sawyer listed the "types" or "modes" of interpreting to illustrate the content validity of interpreting examinations. See 2.3.2.b below.*

Using these sample statements, the following sections will review and discuss the three types of validity in relation to the interpreting assessment.

2.3.2.a External validity

As mentioned above, there are two types of external validity – concurrent and predictive, which denote different times, or the sequence, when a person's validity performance on the external measure is obtained (Salvia & Ysseldyke, 1995: 168). In other words, the test validation by the external measures are done by comparing the results of a test, usually test scores, with some other tests taken by the same test candidates at similar times of the test being validated, i.e. *concurrent*, or some time after the test, i.e. *predictive* (Alderson et al., 1995: 177, 180). Therefore, the external measure itself must be valid so that it can be used to establish the validity of another test on a different occasion or at a different time (Salvia & Ysseldyke, 1995: 168).

Predictive validity

Sawyer's two statements actually refer to the predictive validity of the interpreting examination. They ask if the graduates' test results could allow an accurate estimation of the student's score on an external measure obtained at some time in the future, and of the student's actual readiness to work in the industry (Salvia & Ysseldyke, 1995: 169). The industry-readiness evidence for a final examination's predictive validity can be obtained retrospectively only when interpreter graduates start working in the market, preferably being monitored with a quality assurance scheme like the one proposed by Kalina as discussed in 2.1.5, or being compared with test results from an industry-administered interpreting examination that is for contracting work or in-house staff positions. This in effect is also a validation on the training programme to see if its overall curriculum objectives were met.

For other summative assessments, such as the entry-level tests or in-study placement tests that aim to determine students' ability to study interpreting well in the chosen pathway, predictive validity of the tests is crucial. The evidence of the predictive validity for these tests can be obtained by systematically monitoring and comparing students' examination results over the course of the study, or by asking the teachers their observations and comments on the students' performances in the classroom (Alderson et al., 1995: 182).

Concurrent validity

If Sawyer's second statement is modified as "Interpreter students can pass same or similar tests again shortly after the first test", it conforms better to the definition of the concurrent validity, which in essence "involves the comparison of the test scores with some other measure for the same candidates taken at roughly the same time as the test" (Alderson et al., 1995: 177). However, gathering believable external data is difficult in actual practice (ibid: 178). In situations where a valid parallel version of the same test measure does not exist or difficult to access, therefore, other external measures for obtaining evidence must be considered, such as calling upon the judgements of the candidates' achievement by teachers or professional users (Salvia & Ysseldyke, 1995: 169), or even the candidates' self-assessment of their own abilities to be tested, though care needs to be taken when interpreting such data as evidence for validation (Alderson et al., 1995: 177).

In consideration of the above, there are examples of attempts and efforts to ensure the concurrent validity of the interpreting examinations, such as Sawyer's suggestion to ask "whether successful professional practitioners perform satisfactorily on the final examination in question" (2004: 101), the practice of employing external examiners in the UK's higher institutions to moderate examinations (Liu et al., 2008: 15), and the

practice of holding joint professional interpreting examination between Taiwan's two graduate Translation and Interpretation institutions (ibid: 29).

2.3.2.b Content validity

Content validity asks if a test covers the skills for good performance and all the aspects of the taught subject (Gipps, 1994: 58). It is "the *representativeness* or *sampling adequacy* of the content – the substance, the matter, the topics – of a measuring instrument" (Kerlinger, 1973: 458). Conference interpreters are expected to possess the ability to work in various interpreting modes, such as those in Sawyer's four sample test items, i.e. the four modes of interpreting – consecutive interpreting, liaison interpreting, and simultaneous interpreting with and without text. Therefore, they are examples of such test items that can represent the domain of interpreting skills to be measured.

Other sources to consider the evidence for content validity include "the degree of authenticity of the subject matter and terminology being tested" (Sawyer, 2004: 99). The test designers of interpreting examinations need to consider not only the difficulty level of speech, but also subject areas or speech topics that a conference interpreter is more likely to encounter in real-life interpreting situations (ibid: 100).

Validation of test content

But who determines what is to be used as test items, and how? When assessing students, test designers and examiners must have a clear understanding of the traits, abilities, and skills that are to be measured. Test developers typically rely on panels of experts for judgments about the appropriateness of test content (Alderson et al., 1995: 173; Gipps, 1994: 59; Salvia & Ysseldyke, 1995: 163). In other words, the relevance of the test items to the content of a particular domain tends to be based on the experts' professional experience and judgement.

However, it has been found that experts often disagree with each other (Alderson et al., 1995: 175; Sawyer, 2004: 188) so the content validation needs to be done in a systematic way, such as using some data collection instrument for the experts to make and record their judgements. The common approach is firstly to gather expert opinions on the content to be tested, which are compiled and sorted for further considerations by professional judges, and then based on the resulting judgements an estimate of the test's content validity is derived (Alderson et al., 1995: 173-175).

When defining the domains for the content validity of a test, there are three considerations: "the appropriateness of the types of items included in a test, the completeness of the item sample, and the way in which the items assess the content" (Salvia & Ysseldyke, 1995: 163). Sawyer referred these considerations as "coverage of the domain" and "aspects of test content" (2004:100). These guidelines can help to define the content to be measured better and to gather evidence for test validation.

- *Aspects of test content*

The aspects of test content refer to how the test items assess the content. The measurement of the test content is multifaceted, which includes, for example, "the themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring" (APA Standards, 1999: 11; in Sawyer 2004: 99).

As the method of measurement may affect the outcome, this aspect of validity may depend on the favoured responses types, and one consensus is that the assessment methods should closely parallel those used in instruction (Salvia & Ysseldyke, 1995: 165). In language testing, Alderson et al. refer to this aspect of test content validity as *face* validity and *response* validity. The former mainly concerns people's holistic judgement to see if the test looks like something one might do in the real world

(Alderson et al., 1995: 172); whereas the latter aims to ascertain how test takers respond to test items because "[t]he processes they go through, the reasoning they engage in when responding, are important indications of what the test is testing" (ibid: 176).

The aspects of test content are usually described in the *test specifications*, which is an important record for validation (Luoma, 2004: 117). More discussion of test specifications will follow later in 2.4.

- *Coverage of domain*

The coverage of the domain considers the appropriateness and the completeness of test items. Incomplete assessment of a domain will often result in an invalid appraisal (Salvia & Ysseldyke, 1995: 165). For example, a test cannot be valid if it relies on reading tasks to assess the test takers' interpreting ability because the test tasks are not appropriate to the ability being measured. Also, if a test programme for assessing conference interpreters does not include a test item of simultaneous interpreting, the validity of its results cannot be ascertained because one of the important work modes of conference interpreters is not tested.

In addition, conference interpreters need to work in a vast range of different subject areas, or domains, in an international arena. For content validity, the speech texts chosen for the examination tasks are often selected from those that conference interpreters may actually encounter in the work place, i.e. the *face* validity. The usual choices of topics when setting the examination tasks may range from politics, to business, to economics, to technology and current affairs.

Of course, it is impossible and impractical to include everything in a single interpreting examination. As Sawyer said,

> a debate on whether all interpretation is the same is defeatist; that is, there is no point in debating, for the purposes of assessment, whether a day-long

> simultaneous conference on wood processing is equivalent to an emergency doctor-patient telephone call due to the allergic reaction of a child. [Therefore,] an alternative approach would be to identify and meet the needs of a given setting.
>
> (Sawyer, 2004: 116)

Therefore, depending on the educational goals of the curriculum and the purposes of the assessment for interpreting, representative subject domains should be identified for a more focused assessment of interpreting skills and abilities in various settings. The approach of adopting the idea of subject domains for interpreting examinations is consistent with the multi-perspective and situational considerations in quality assurance for interpreting discussed earlier in 2.1. Again, empirical studies are required to validate the test designs regarding the identification of suitable subject domains and the underlying interpreting skills and abilities to be assessed, i.e. the *test construct*.

2.3.2.c Construct validity

Construct refers to, "a theoretical conceptualisation about an aspect of human behaviour that cannot be measured or observed directly" (Ebel & Frisbie, 1991: 108; in Alderson et al., 1995). Construct validity, therefore, concerns the underlying skill being assessed that is indirectly reflected in a test performance (Gipps, 1994: 58).

In the case of interpreting, as the nature of quality assurance for interpreting is multidimensional (2.1), the domain of interpreting is, as Hoffman put it, "sensitive to the audience and speaker and their relations and goals, sensitive to world knowledge and context as well as topic, and sensitive to status relations, loyalty shifting, and nuance as well as to literal meaning" (Hoffman, 1997: 204; in Sawyer, 2004: 116). To cover all these would involve a complex set of skills and a multiplicity of test constructs to be measured. Thus, it is no longer "interpreting" *per se* that is being tested, but a wide range of communication abilities in cross-cultural and cross-lingual situations, which involve the interpreters' cognitive, linguistic and social skills with the support of

knowledge of the relevant subject matter. In 2.2.1, for example, the four requirements for interpreters: *mental model, knowledge base*, *linguistic expertise*, and *interpreting skills* may be regarded as the working constructs of interpreters' ability to interpret. As these constructs cannot be measured directly, they need to be demonstrated and explained by test performance based on clearly defined assessment criteria, which is what the present study tries to find out.

Verifying construct validity for interpreting assessment

"[C]onstructs cannot be developed and measured in isolation from one another, but must be part of an integrated viewpoint" (Sawyer, 2004: 98). The integrated viewpoint is usually grounded on a single theoretical framework or a combination of several that provide concepts and terminology, for example, to help us explain what we mean by good interpreting performances. The professional judgement in the field, and the educational philosophy and objectives of the training institutions are "influential factors in the discussion to reach a consensus on construct definitions" for the interpreting assessment (ibid: 99).

The evidence for construct and content validity is substantially overlapped (Messick, 1988: 38; in Sawyer, 2004: 99). Alderson et al. explained that

> In effect, this form of construct validation proceeds much as does content validation: experts are selected, given some definition of the underlying theory, and asked to make judgements after an inspection of the test as to its construct validity (1995: 183).

For interpreting assessment, useful theoretical framework may be derived from professional experience and research work that test writers can rely on for the identification of test constructs, and the assessment criteria to measure them. For

example, the findings from the user-survey studies (2.1.1) and subsequent studies that shed light on the assessment criteria for interpreting performances, including, among others, Pöchhacker's model of quality standard for interpreting (2.1.4) and Kalina's integrated QA framework for interpreting (2.1.5).

At the beginning of 2.3.2, sample definitions from Sawyer were presented as example constructs for interpreting examinations, which are recapitulated here for discussion. In an interpreting examination, the examinees should demonstrate that they have the ability to (1) interpret with faithfulness to the meaning and intent of the original, (2) use appropriate language and expression, (3) apply world knowledge and knowledge of subject matter, and (4) demonstrate acceptable platform skills and resilience to stress (Sawyer, 2004: 97).

These four construct statements relate closely to Pöchhacker's model of quality standard for interpreting with *accurate rendition* at the core, followed by *adequate expression, equivalent effect*, and *successful communicative interaction* (see Figure 2-2). Sawyer's construct statements elaborate how the four dimensions of quality standards can be observed in the context of an interpreting examination. For example, the *accurate rendition* needs to be observed by comparing the interpretation output with the "meaning and intent of the original", the *adequate expression* can be observed by the examinee's use of "appropriate language and expression", and so on.

At operational level, however, the relevant interpreting abilities, i.e. the test constructs of interpreting, still have not been precisely defined, and the assessment criteria have not been developed in coherence with the constructs, either (Sawyer, 2004: 98). For instance, professional organisations like the AIIC and the Directorate General for Interpretation[6], base their assessment criteria for selecting members and evaluating

[6] The Directorate General for Interpretation (formerly known as SCIC) is the European Commission's interpreting service and conference organiser.

novice staff interpreters on the same source of professional judgement and empirical data. Peng (2006:19) listed some samples of the professional organisations' criteria for assessing simultaneous interpreting as below:

1) Rigour and consistency
2) Faithfulness to original (substance and style)
3) Quality of communication with audience
4) Calm, regular delivery
5) Avoid literal/word for word translation
6) Correct, spontaneous use of TL

Assessment criteria like the above "lack clear definition and proper organisation" and are "not sufficiently clear" (Hartley, Mason, Peng, & Perez, 2004: 3; Peng, 2006: 18), and are thus open to discussion and subjective judgement.

Clearly defined assessment criteria come from clearly defined test constructs. Difficulties will arise when ill-defined constructs are put into a testing context and it is made operational. Take Sawyer's construct statements for example, when we say "the ability to interpret with faithfulness to the meaning and intent of the original", what assessment criteria can we use to measure this construct so that it is clear and acceptable to every examiner who may sit on any given examination panel? When we say "use appropriate language and expression", what does it mean by saying appropriate? Appropriate to whom under what situation? The above six assessment criteria do not give more explanations than Sawyer's construct statements for the examiners to assess the interpreting performances; they are open to subjective judgement.

This then relates to the concerns about professional judgement as discussed in 2.3.2.b. Alderson et al. emphasised that in this construct validation approach, i.e. using professionally-judged theoretical framework, "the theory itself is not called into question: it is taken for granted" (1995: 183). Therefore, if the theory itself is not firmly grounded, the construct validation cannot be intrinsic. Empirical studies are required to

obtain data as evidence to substantiate the professionally-derived theories for the interpreting assessment. Hence, the endeavours of the present research study.

Being a performance-based assessment, the main issue lies in the fact that many subjective factors are involved in not only the test validation process, but also the assessment of interpreting performances. A good level of reliability in the examiners' judgements is required to maintain the test validity, which will be reviewed next.

2.3.3 Test reliability

The concept of reliability in testing can be simply defined as "the extent to which an assessment would produce the same, or similar, score on two occasions or if given by two assessors" (Gipps, 1994: vii). Without reliability, a test cannot be considered a valid form of assessment and its results and decisions cannot be considered meaningful and useful. For example, if we use bathroom scale as an analogy, a reliable bathroom scale will give an accurate measure of weight no matter where and when it is used and regardless of who stands on it. Therefore, reliability can be viewed as another factor in testing that is subsumed under the concept of validity. It is usually analysed separately to simplify the discussion of validity (Salvia & Ysseldyke, 1995: 174).

Of course, measuring a skill or educational achievement is much more complex than measuring body weight. In education, students learn and improve, and the aim in testing is to measure the systematic changes as a result of students' improvement. Errors in testing will occur due to unsystematic changes from factors that are not related to students' learning outcomes, such as the physical testing conditions, the administration of examination, or inconsistent marking practice. So "the higher the proportion of systematic variation in the test score, the more reliable the test is. A perfectly reliable test would measure only systematic changes" (Alderson et al., 1995: 187).

When it comes to testing, we are interested in generalising what is observed under examination conditions to other occasions. For example, a student's interpreting performance in an examination needs to be generalised to the real world conditions, such as business meetings or conferences. Otherwise, the examination results would be of little or no value, i.e. poor *face* validity. Generalisation is necessary because it is impractical to assess individuals on everything in the whole domain, so the assessment needs to be based on a sample performance that is representative of the domain. It is similar to drawing inferences in test validation (*c.f.* 2.3.2.b), but generalisation is expected to be made in a consistent way. Therefore, the concept of "generalizability" can be regarded as the link between validity and reliability (Gipps, 1994: 76).

There are three approaches to estimate test reliability by generalisation: (1) generalising to different examiners, (2) generalising to other test items, and (3) generalising test results over time (Salvia & Ysseldyke, 1995: 136).

2.3.3.a Examiner-related reliability

To illustrate examiner-related reliability in a test, let us return to the *Prelude* and consider some of the teacher's questions in his reflection on the interpreting examination:

> *Have the examiners picked up every error in each student's interpreting performance? If not, was this examination fair? If this examination had been assessed by another panel of examiners, would they have given similar marks to those that the examiners gave in this panel?*

Questions like these are concerned with the *inter-rater reliability*[7], which refers to the "agreement between raters on the same assessment task" (Gipps, 1994: 67). When

[7] For the convenience of discussion, three terms – *rater*, *assessor*, and *examiner* – will be used interchangeably in this thesis. They all refer to a human subject who makes judgement and gives a mark to test takers' performances.

assessing students by following a test programme, it is assumed that the same student's performance will receive a similar result from different examiners. If different examiners assess the student differently in a test, we may have less or little confidence in the meaning and usefulness of the student's mark in that test.

Another examiner-related reliability is the *intra-rater reliability*, which refers to "the agreement of the same rater's judgments on different occasions" (ibid: 67). A reliable examiner should be able to apply the same assessment criteria consistently and "give the same score for equivalent test performance on separate occasions" (Sawyer, 2004: 102). Otherwise, it will be difficult to generalise and make sense of the test results even when the test itself is a valid assessment tool.

The examiner-related reliability is most critical to subjective forms of assessment, holistic marking and observation (Salvia & Ysseldyke, 1995: 146). Given the requirement of professional judgement in evaluating interpreter performances, interpreter examinations are inherently subjective in nature and often lack inter-rater agreement (see 2.3.2.b). Therefore, "one of the most relevant types of reliability for interpreter assessment is the consistency of scoring across raters" (Sawyer, 2004: 102-103). Although it would not be realistic to expect all examiners to judge in the same way all the time, a high degree of overall consistency is essential for the test to be considered reliable by its users (Alderson et al., 1995: 129). Therefore, a uniform basis for marking needs to be established to assist and monitor the examiners' judgement; this will be discussed later in 2.4.

2.3.3.b Test item-related reliability – internal consistency

The second type of reliability concerns the *internal-consistency* of a test and is an issue involved in test item development (Campbell & Hale 2003: 220). In respect to test item-related reliability, it is expected that "similar but different test questions would

give us similar results; we would like to be able to generalise to other similar test items" (Salvia & Ysseldyke, 1995: 136). The similar tests are so called *alternate forms* of a test, which measure the same trait or skill to the same extent on the same population of test takers. Therefore, several names are used when referring to this type of reliability, such as *alternate form* reliability, *equivalent form*, or *parallel form* reliability (Campbell & Hale, 2003: 140; Gipps, 1994: 67).

Internal consistency of interpreting examinations

One method often used for estimating a test's internal consistency is the *split-half* method, i.e. splitting a test into two halves to see if the scores on these two halves are consistent. However, the split-half method does not work well with interpreting and translation examination tasks because when a speech or a text is divided into two parts, it is impossible to keep the two parts independent (Campbell & Hale, 2003: 220). For example, the first half of a speech may contain the opening remark and the second half the conclusion, thus the two halves in effect will test different content in an interpreting examination, and their results cannot be compared. It would be like chopping a bathroom scale in half and still expecting the two halves to work independently.

Therefore, independent and reliable equivalent forms of test items need to be developed to maintain the internal consistency of interpreting examinations. The approach is to check whether speeches used as examination material vary in difficulty between examinations that are supposed to measure the same level of interpreting skills (Sawyer, 2004: 101-102).

However, it is not always easy to produce equivalent forms of a test, especially when it comes to performance assessments like speaking and writing tests (Gipps, 1994: 67-69). Campbell & Hale also found that determining the degree of difficulty of the source material for both interpreting and translation tests remained a major barrier to

improving test reliability because there was no objective method for determining the difficulty of source materials. The situation was further compounded by the need for the regular introduction of fresh examination materials into the interpreting and translation testing regime. The lack of an objective mechanism to determine the difficulty of the examination tasks would potentially generate highly unreliable marks (Campbell & Hale, 2003: 219). To tackle this problem, some interpreter educational institutions have developed guidelines for selecting assessment materials to help reduce the subjective element, making the test items more equivalent to each other (Liu et al., 2008: 6-11). Test design issues will be reviewed and discussed further in 2.4.

2.3.3.c Test stability – reliability when generalising over time

The third type of reliability concerns the *stability* of the test over a period of time. In the educational context, "we would also like to assume that the behaviour we see today would be seen tomorrow (or next week) if we were to test again" (Salvia & Ysseldyke, 1995: 136). For example, it is unlikely that a novice student interpreter can master simultaneous interpreting overnight or in a week and start working at international conferences. So if the student is tested today and tested again tomorrow using a reliable test of equivalent form that measures the same level of interpreting skills, the test results should be very similar. Although over a longer period of time interpreting skills can be improved and become more noticeable, over short periods of time the development of the interpreting skills should be relatively stable. Therefore, examination tasks for assessing educational traits with a slow developmental pace "must produce sufficiently consistent and stable results if those results are to have practical meaning for making educational decisions" (ibid: 139-140).

However, it is impractical to use the test-retest method to estimate the stability of interpreting examinations because the student interpreters would have been familiarised

with the test materials (Sawyer, 2004: 102). Therefore, the first two types of measures for generalisation are important for the interpreting assessment, especially the examiner-related reliability as discussed previously.

2.3.4 Tensions between validity and reliability

Nevertheless, "assessment is not an exact science" (Gipps, 1994: 167); it would be misleading to suggest that we could really "accurately" assess the acquiring of complex skills in performance-based assessment (ibid: 70-71). This view emerged mainly due to a tension between validity and reliability. Understanding this tension may help to identify solutions to the problems and avoid pitfalls when developing the interpreting examinations.

The tension emerged because of a paradigm shift in assessment in the past few decades from psychometrics to a broader model of educational assessment (Gipps, 1994: 1-17). Despite the importance of validity, test development in psychological and standardised testing tended to emphasise reliability, and test validity was often sacrificed when attempting to achieve highly accurate and replicable testing. Therefore, performance-based assessments were developed as an attempt to redress the balance between reliability and validity (ibid: 76).

Assessments based on the traditional psychometric test theory are norm-referenced and their reliability measures rely on statistical correlation techniques. A norm-referenced assessment emphasises the differences among individual students, and the test items aim to separate the students' ability levels and rank them by assigning different marks. Then the marks are subjected to statistical correlation calculations to estimate the test reliability.

However, statistical correlation techniques do not work well with criterion-referenced assessments, e.g. performance assessments. For a performance-based assessment, the crucial aspects of the test validity and reliability are the quality of performance and the fairness of scoring, not the test's ability to be replicated and the generalisability of the performance (Frederiksen & Collins, 1989; Moss, 1992: 250); in Gipps, 1994: 172). As discussed previously, it is difficult to maintain high measures of internal consistency in performance-based assessments because they rely on professional judgement and is subjective in nature. Performance assessments, therefore, depend on certain criteria to help produce reliable test results. This triggers the tension because the test designs differ between norm-referenced assessment and criterion-referenced assessment.

The design of criterion- referenced assessments does not emphasise the differences among individuals, but aims to assess the examination performances based on a set of criteria or professional standards (i.e. construct and content validity), such as those in the interpreting and translation examinations. So, for instance, if most performances in an examination meet the defined assessment criteria, the range of marks may be narrow. Therefore, it would be misleading if correlation statistics, which assume high levels of discriminative power of the test items, were used for estimating reliability for criterion-referenced assessment (Gipps, 1994: 68).

A balanced view

Although tensions exist, reliability and validity are important dimensions of test development. They remain essential quality assurance devices for assessments. Gipps suggested rethinking validity as "a question of prioritizing and specifying the responsibility of test developer, policy-maker and user" (ibid: 171), and for the test to be useful to those concerned, the assessed constructs and appropriate test use must be

clearly articulated so the test results can be contextualised (ibid: 170-172). So that "the highest validity and optimum consistency and comparability for a particular purpose" of a test can be achieved (ibid: 173). In language testing, Alderson et al. explained the balancing act between the two as below:

> In practice, neither reliability nor validity are absolutes: there are degrees of both, and it is commonplace to speak of a trade-off between the two – you maximize one at the expense of the other. Which you choose to maximize will depend upon the test's purpose and the consequences for individuals of gaining an inaccurate result (1995: 187).

In other words, the balance between validity and reliability (i.e. optimum consistency and comparability), can and should be adjusted according to the purpose of the test.

For example, if the purpose of an interpreting examination is to determine an interpreter's ability to work for a specific international organisation, the test content coverage should be the main consideration to achieve the maximum test validity. For fairness, optimum consistency of the test items should be applied to all examinees. The crucial aspect of test validity and reliability, therefore, is the quality of the interpreting performance and the fairness of scoring in relation to the specific organisation. This test and its results may not need to be replicated and generalised for other organisations or businesses because of the contextualised purpose of the interpreting examination.

However, when an interpreting examination is for testing interpreting skills in general that can be applied to a wider range of educational purposes like formative assessments, the content validity needs to be adjusted and the comparability of the test items need to be considered according to the purpose and objective of the educational curriculum. Comparability of test programmes can be:

> achieved through *consistency of approach* to the assessment by teachers; *a common understanding of assessment criteria*; and that performance is evaluated

> fairly, that is, according to the same rubric by all markers. These can be achieved by a combination of training, moderation and provision of exemplars.
> (Gipps, 1994: 174).

Recognising these tensions and the approaches to finding a balanced view between validity and reliability in assessment, it is clear that the attention needs to rest on the examiners and the examination procedures in performance-based examinations like interpreting assessment. Interpreting examinations need to be given clearly defined contexts for their usage, and require a consistent approach to the developing and marking of the assessment tasks (Gipps, 1994: 103-105; Campbell & Hale, 2003: 221; Sawyer, 2004: 102).

2.3.5 A challenge – validating interpreting examinations

Section 2.3 has reviewed and discussed the concepts of test validity and reliability and their implications for interpreting examinations. Validity is the cornerstone of any assessment to be used in a meaningful way. A test regime can be validated by looking at its content and constructs, which need to be based on empirical data from relevant professional domains. They need to be clearly defined to be operational in the actual testing process to achieve reliability. The concept of test reliability is an extension of validity. A test without an acceptable level of consistency cannot be considered valid. A test's reliability is determined by looking at three areas of generalisability: examiners, test items, and test stability. The tensions between validity and reliability have also been discussed, especially those between norm-referenced and criterion-referenced assessments. A balanced view may ease the tensions by clearly articulating the context of the test use and optimising the testing procedures so that comparable and dependable tests can be purposefully designed for specific applications.

From the reviews and discussions above, professional judgement appears to be the crucial element that determines the test validity and reliability of interpreting examinations. However, "as often stated in the literature on expertise, experts often disagree" (Sawyer, 2004: 188). Sawyer's case study shows that the expertise of interpreter examiners does not necessarily yield agreement in the exercise of their professional judgement and "extreme fluctuation in professional judgment is evident" (ibid). More systematic studies on the assessment procedures of interpreting examinations were urgently called for (ibid: 187-189).

As the social consequences of interpreting examinations become more evident, interpreting examinations have to be more carefully scrutinised. It was found that there were "strong interest", "openness, willingness, and dedication" in the community of interpreters in improving interpreting examinations (ibid: 188). The survey studies of quality assurance in interpreting (2.1) accumulated a substantial amount of knowledge that may serve as a foundation for the content validation of interpreting examinations. However, empirical studies on the guidance to help reduce the subjectivity in the interpreting examinations need to be conducted.

As outlined in 1.3, one important research topic this research study intends to understand is how the examiners exercise their judgement *during* the interpreting examinations. This study area may lead to the creation of a standardised foundation, such as a conceptual model of the interpreting examination, for understanding and explaining how to assess the multidimensional, themed communication act of interpreting in a valid and reliable manner.

In 2.4 below, existing knowledge from the language testing discipline will be further explored and some approaches for test design and development will be reviewed in relation to the improvement of interpreting examinations.

2.4 Test design and development for interpreting examinations

The test design and assessment criteria for evaluating interpreters' performance have been described as "intuitive" (Campbell & Hale, 2003: 211). Liu et al. also found that many interpreter educational institutions relied on the judgement and experience of the staff members to design, administer and mark interpreting examinations, but often with no basis of empirical studies for the test items and test procedures (2008: 35).

The evaluation of the appropriateness of a test and its application does require professional judgement for validity reasons. However, professional judgement alone is not a sufficient basis for decision-making. There should be some protection against unsound professional judgment (Messick, 1989; in Sawyer, 2004: 104). Professional judgement "should be wielded with considerable care and circumspection" by using empirical data to reduce subjectivity when selecting test content and developing assessment criteria (Sawyer, 2004: 104). Many interpreter educational institutions recognised the need to reduce the risk of subjective judgement and put in place guidelines for setting the difficulty level of the examination tasks and the marking criteria. Nevertheless, they found it difficult to follow the guidelines because of the need to retain the authenticity of the task in the performance-based assessment, and because of the limitation of live marking practice (Liu et al., 2008). A uniform and practical basis is needed for developing the examination tasks and marking the performances, i.e. a standardised procedure for interpreting examinations.

Being a performance assessment, the design and development of interpreting examinations may benefit considerably from the experiences of the disciplines in educational assessment and language testing (Campbell & Hale, 2003: 221). Interpreting examinations are more similar to language tests, such as speaking tests,

than to other types of assessment, such as writing an essay. Both interpreting and speaking tests, for example, are examinations that require the examinees to respond by using spoken language, and the examiners' judgements are also both subjective in nature. Therefore, the sections below will look at the test design approaches from language tests as a basic framework for discussions of interpreting test design.

2.4.1 Test specifications – a framework from language testing

In language testing, the development of a test usually begins by specifying the various facets of the assessment method. The test designer, be it a language teacher or an interpreter trainer, should have ideas about what to test, how to test, and what the marking criteria should be. These ideas are put in a written document called a *test specification* as guidelines for the development of the test (Luoma, 2004: 113). Alderson et al. described the document as follows.

> A test's specifications provide the official statement about what the test tests and how it tests it. The specifications are the blueprint to be followed by test and item writers, and they are also essential in the establishment of the test's construct validity (1995: 9).

Alderson et al. also distinguish between the documents *test specifications* and *user specifications* (or *syllabus*). The former is more detailed and is often for internal use only, such as for test developers; the latter is a simplified version of the test specifications and is a public document for the test users to know what the test will contain (ibid). Much language testing literature has provided suggestions on how the test specifications should be written in various formats and frameworks (Alderson et al., 1995; Bachman and Palmer, 1996; Luoma, 2004; Sawyer, 2004). No matter which

format is selected, the document generally contains the information as summarised by Luoma below:

> The specifications contain the developers' definitions of the tasks and rating criteria to guide the development of comparable tasks and the delivery of fair ratings. The specifications record the rationale for why the assessment focuses on certain constructs and how the tasks and criteria operationlise them (2004: 113).

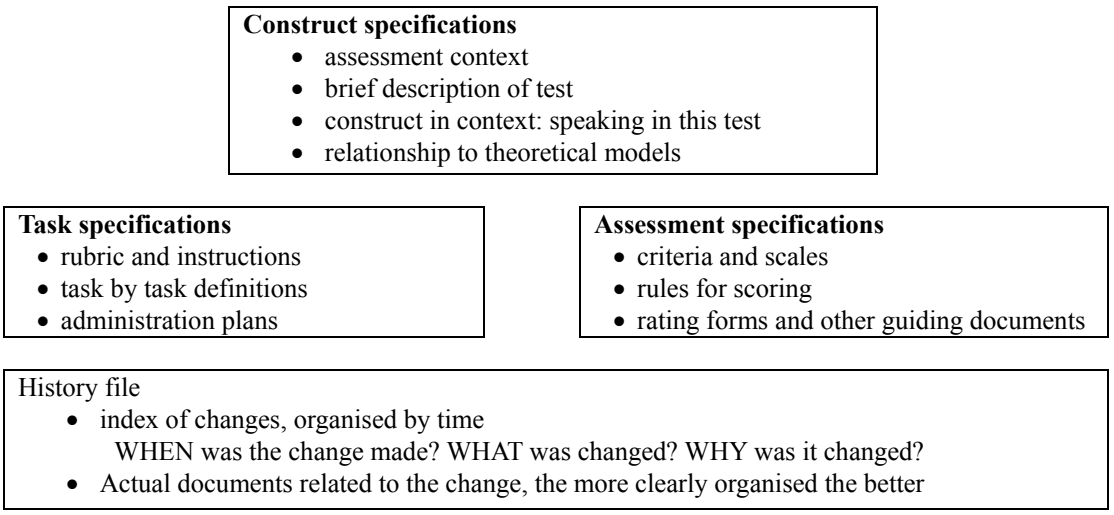
Alderson et al. provided a detailed check list, shown below, for the content of language test specifications that can be tailored for different needs and test uses.

> The test's purpose
> Description of the test taker
> Test level
> Construct (theoretical framework for test)
> Description of suitable language course or textbook
> Number of sections/papers
> Time for each section/paper
> Weighting for each section/paper
> Target language situation
> Text-types
> Text length
> Language skills to be tested
> Test tasks
> Test methods
> Rubrics
> Criteria for marking
> Descriptions of typical performance at each level
> Description of what candidates at each level can do in the real world
> Sample papers
> Samples of students' performance on tasks
>
> (Alderson et al.,1995: 38)

The check list covers much information about test designing. Depending on the purpose and complexity of the examinations, different authors have suggested slightly different formats and structures for writing the test specifications document.

When discussing the test specifications for speaking tests, Luoma (2004) proposed a modular structure of specifications as shown in Figure 2-4. In this modular structure, the specifications are seen as a single document that consists of three modules: construct, task, assessment, plus an optional history file. The advantage of the modular structure is that each module focuses on one conceptual part of the test development; the modules then can fit together coherently as an important document for test validation.

Figure 2-4 Modular structure of test specifications



Source: (Luoma, 2004: 117)

The main aspect in the *task* specifications is about the test interface for the examinees, which allows them to perform the examination tasks in the optimum conditions so that their performances can be measured in a way that is consistent and reliable. The *assessment* specifications describe how the examiners should assess the examinees' performances by using various test instruments, such as assessment criteria and rating scales, and following the marking guidelines. These two specification modules then are linked and unified under the *construct* specifications module so the test design is coherent as a whole. As for the optional history file, it is a record of the test development, providing useful information for monitoring and sharing the experiences accumulated in the course of the test design (ibid: 115-118).

Documenting as many of the facets of test design and use as possible, i.e. producing the test specifications, will help build a case for the validity and reliability of test items in interpreting examinations (Sawyer, 2004: 107). The documentation of the test specifications should be an on-going process of test development, an iterative process that requires trial and error over an extended period of time so that the final test product can be presented as explicitly as possible.

Ideally, the production process of the test specifications should involve those who are interested in the same field discussing and exchanging opinions. Thus, different viewpoints regarding the purpose, nature and the format of the test may be reconciled in the design and development process. The contradictory needs for standardization and authenticity of the performance assessment task, therefore, can also be balanced during the reconciliation process (ibid: 108). In other words, the exercise of producing the test specifications may also help in finding a balanced position in relation to test validity and reliability as discussed in 2.3.4.

2.4.2 Development of test specifications for interpreting examinations

Using the modular structure of test specifications as a basic framework, the following sections review and discuss the state of test design of simultaneous interpreting examinations.

2.4.2.a Construct specifications

Construct validity in interpreting examinations was reviewed earlier in 2.3.2.c. Sawyer's four sample construct statements (2.3.2) are a good starting point for writing the construct specifications. The main problem of current construct statements for many interpreting examinations is the lack of definiteness and operational practicality

(2.3.2.c). For example, China Accreditation Test for Translators and Interpreters (CATTI) published an Examination Outline of Level 2 simultaneous interpreting[8]; the Outline lists four basic requirements and qualifications for the examinees, which could be regarded as the construct statements of the examination. The four statements are extracted and translated into English as below:

1. Solid foundation ability in fluent usage of both Chinese and English languages.
2. Broad background knowledge, such as in politics, economics and cultures.
3. Sophistication in the skills of simultaneous interpreting.
4. Good psychological quality and resourcefulness.

They are similar to Sawyer's construct statements and both include the four basic constructs as identified from Kalina's (2005) integrated framework for interpreting quality assurance: *language competence*, *interpreting skills*, *background knowledge*, and *personal aptitude*. (see 2.1.5 and 2.3.2.c).

Nevertheless, these construct statements need more precision parameters for the operational practicality of the interpreting examinations. For example, what are the skills of simultaneous interpreting? How can they be described and measured? What is a good psychological quality for conference interpreters? How can it be identified in an interpreting examination? As for language ability and background knowledge, how can they be described and assessed, i.e. what assessment criteria can be used to measure these test constructs? All these questions require clarifications.

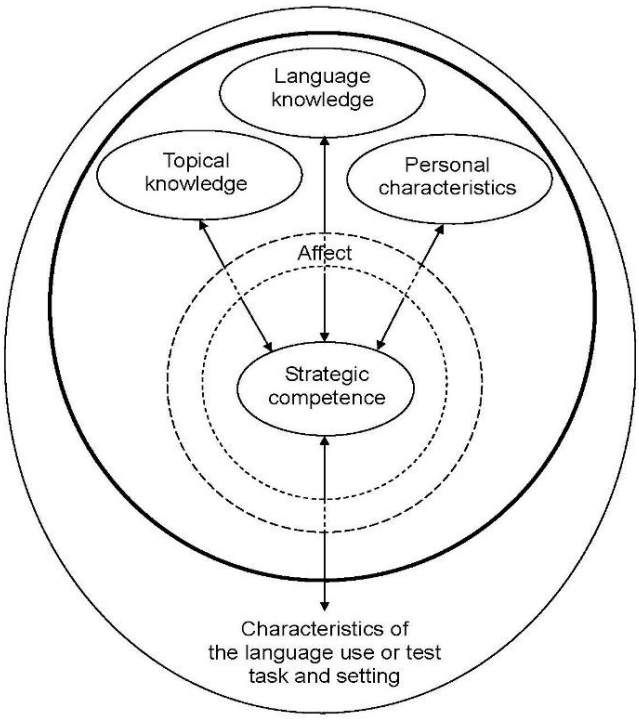
It is also important to identify a theoretical framework that can coherently present these constructs for interpreting assessment, i.e. explaining their relations. As mentioned before, the knowledge in the discipline of language testing is a good source to assist the

[8] The Examination Outline for Level 2 simultaneous interpreting of CATTI is available online at http://bbs.catti.china.com.cn/down/ syllabus _EN_SI2.pdf. Accessed 17 May 2009.
CATTI source text in Chinese: 1. 基本功扎实，熟练运用中英文两种语言。2. 知识面广，有比较宽泛的政治经济文化等背景知识。3. 熟练运用同传技能。4. 具备良好的心理素质和应变能力。

development of interpreting assessment. Below we shall review a language model to see how the construct of language can be explained and described, and use it as a theoretical framework for reference in the context of interpreting assessment, which is also closely related to the use of languages.

Describing the language construct in language testing

Figure 2-5 Some components of language use and language test performance



Source: (Bachman and Palmer, 1996: 63)

Based on theories about communicative competence and empirical results from a multitrait-multimethod study, Bachman and Palmer (1996) proposed a model for describing the language users, or potential test takers; the model explains the relationships between five hypothesised components of language use and language test performance (Luoma, 2004: 99). The model was conceived as a conceptual basis for organising the thinking about the language test development process (Bachman & Palmer, 1996: 62-63), and was illustrated in a figure as reproduced here in Figure 2-5.

As shown in Figure 2-5, the characteristics of individual language users are depicted with five components i.e. constructs: *language knowledge*, *topical knowledge*, *personal characteristics*, *strategic competence*, and *affective factors*; the double-headed arrows indicate the interactions or relations between the components. Among these components, *strategic competence* and *affective factors* serve as mediators of the other three components. *Strategic competence* was conceived "as a set of metacognititive components, or strategies, […] that provide a cognitive management function in language use" (Bachman & Palmer, 1996: 70); whereas *affective factors* refers to a schemata of the language users in terms of their past emotional experiences in similar contexts (ibid: 65).

This model of language use is goal-driven: *strategic competence* enables the language users to set goals and utilise the *language knowledge* and *topical knowledge*, so that they can interact with the settings in which the language use takes place, e.g. a test task, taking into consideration the language users' *personal characteristics* and *affective factors*. For a detailed explication of this model, see (Bachman & Palmer, 1996: 61-82), and (Luoma, 2004: 97-101).

Based on the model of language use above, Fulcher further developed a framework for describing the test constructs of second language speaking (2003: 18-49). Fulcher suggested that contextual factors should be considered in the construct definition and that "test purpose should drive the definition of the construct, its range and generalisability" (ibid: 19). Any construct of *speaking* is multifaceted in view of the richness of what happens in a process as complex as human communication (ibid: 25). This multifaceted view in language use is not dissimilar to the multidimensional perspective in assessing the quality of interpreting. A summary list of items that may be included in the construct definitions of a speaking test is shown in Table 2-1.

Table 2-1 A framework for describing the speaking construct

Language competence	**Strategic capacity**
Phonology • Pronunciation • Stress • Intonation Accuracy • Syntax • Vocabulary • Cohesion Fluency • Hesitation • Repetition • Re-selecting inappropriate words • Re-structuring sentences • Cohesion	Achievement strategies • Overgeneralization • Paraphrase • Word coinage • Restructuring • Cooperative strategies • Code switching • Non-linguistic strategies Avoidance strategies • Formal avoidance • Functional avoidance

Textual knowledge	**Pragmatic knowledge**	**Sociolinguistic knowledge**
The structure of talk • Turn taking • Adjacency pairs • Openings and closings	• Appropriacy • Implicature • Expressing being (e.g. refer to oneself or the others)	• Situational • Topical • Cultural

Source: (Fulcher, 2003: 48)

Compared with those in the interpreting examinations, these five groups of items for speaking test give more precise descriptions for the constructs and assessment criteria, which are supported with the language use model to explain their relations. To solve the problems in defining the constructs of interpreting, we can learn from the experiences in language testing, such as the knowledge like these, as they provide references to the use of terminology and framework in describing the construct of speaking.

In the case of interpreting assessment, therefore, the solutions need to be based on empirical studies to obtain data for clarification. Putting the constructs in context with descriptions of the test items or giving an assessment context may help with the clarifications. Theoretical quality models for interpreting, such as Pöchhacker's (2001) model of the quality standards for interpreting (2.1.4) and Kalina's (2005) integrated quality framework (2.1.5), are useful to help clarify the multidimensional and multi-perspective characteristics of quality in interpreting. Nevertheless, their models

lack the structure to coherently integrate and explicate the relations between various ability requirements for performing the tasks of interpreting. Therefore, the job at hand for the interpreting assessment, such as in the case of the present study, is to integrate the existing knowledge of interpreting quality models with empirical data, i.e. to find out how the examiners assess simultaneous interpreting, and put the findings in a wider research context of the interpreting assessment as foundation knowledge for future development and improvement.

2.4.2.b Task specifications

This section reviews the current state of test task design of the interpreting examinations, referring to the knowledge in language testing as introduced in the previous section 2.4.1. As the method of measurement will affect the outcome, it is important to document the test facets with detail in order to maintain the test's internal-consistency (2.3.3.c). The task specifications define the examination tasks, including the task type and the skills to be assessed. The information and instructions about the task also need to be specified (Luoma, 2004: 117), such as "the themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring" (Sawyer, 2004: 99).

Liu et al.'s survey reported that the tasks for interpreting assessment are conducted either by using a recorded speech or by using a live speaker in the examination. Due to the consideration of consistency, speech recordings are adopted by ten of the eleven surveyed institutions[9]. Only two of the surveyed interpreter education institutions

[9] The eleven surveyed institutions are: Monterey Institute of International Studies (USA), Shanghai International Studies University (China), Newcastle University (UK), University of Westminster (UK), Fu Jen Catholic University (Taiwan), National Taiwan Normal University (Taiwan), National Changhua University of Education (Taiwan), Chang Jung Christian University (Taiwan), National Kaohsiung First University of Science and Technology (Taiwan), Leader University (Taiwan), Wenzao College of Languages (Taiwan). Only Wenzo reported that it does not use recordings for interpreting examinations.

(Monterey Institute of International Studies and Shanghai International Studies University) have formal guidelines for setting the test items of their interpreting examinations, whereas the others either have general "principles" or mainly rely on the professional judgement of the interpreter teachers to set the examination tasks (Liu et al., 2008: 6-7, 24-25). In general, these guidelines and principles specify the subject areas, speech types, inclusion of specialised terms, density of information in the speech, and difficulty level of the speech (ibid: 8-10, 26-27). The interpretation of these general guidelines and their implementation among the surveyed institutions are still dependant on the professional judgement of the test designers, who are usually the interpreter teachers (ibid: 10).

Liu et al. also found that the administration of interpreting examinations varies from one institution to another; there are variation in areas such as the instructions to the examination tasks, the preparation time and materials on site (ibid: 27). For most interpreting tasks, details of specific organisations, work context, or themes are often not clearly specified (ibid: 8-10), which will affect the way the test candidates interpret in the examination. In such situation, the *response* validity (see 2.3.2.b) is threatened because the interpreter examinees may not work in a way that can be generalised to what happens in real life.

All the above will inevitably increase the complexity of the test design. With more standardised task specifications, the interpreting examinations may be administered more consistently to ensure the test reliability and validity.

2.4.2.c Assessment specifications

In language testing, assessment specifications describe how the examiners should assess the examinees' performances using test instruments such as assessment criteria and rating scales. The examiners make qualitative judgements on the examinees'

performances, and use rating scales to express or indicate their judgements, usually in the form of numbers, such as a mark, or in the form of categories such as "excellent" or "fair", and so on (Luoma, 2004: 59; Pollitt & Murray, 1996: 74).

As both speaking and interpreting tests observe the spoken language output of the examination candidates, rating scales for speaking tests are useful references for interpreting assessment. Therefore, some types of rating scales in language testing are reviewed here for the discussion on the rating scales for interpreting assessment.

Types of rating scales for language proficiency tests

Traditionally, *global scales* have been used for assessing the language ability. The global scales are "based on the view that language ability is a single unitary ability" (Bachman & Palmer, 1996: 208). As they use a single score to express an overall impression of an examinee's ability, they are also called *holistic scales*. This type of scale is practical and flexible because the scales "allow many different combinations of strengths and weaknesses within a level" (Luoma, 2004: 61-62).

However, holistic scales have also been criticised as having problems, such as the difficulty in making inferences from a single score on the complex components of language proficiency, and therefore, making it difficult for raters to assign levels. Also, the implicit components in a global scale may be weighted differently by different raters in arriving at their single rating (Bachman & Palmer, 1996: 209-210).

Other types of scale are the *analytic scale*, which gives descriptions of the language ability levels. Analytic scales contain assessment criteria and descriptors at different levels of the scale, giving detailed guidance to the raters and provide rich information on specific strengths and weaknesses in examinee performances (Luoma 2004: 67-68). Thus, analytic scales may reflect what raters actually do when rating the test performances, and provide profiles of the examinee's performances being rated,

such as good pronunciation and vocabulary use, but poor grammar when talking, and so on (Bachman & Palmer, 1996: 211).

A simplified version of the analytic scale is the *numerical rating scale*, which contains only the titles or labels of the assessment criteria with a numerical scale attached to each criterion (Luoma, 2004: 76). The numerical rating scale is easy to use, but the interpretations of the marks may vary across the raters because numbers alone are vague (ibid) and, therefore, open to subjective interpretations. More detailed descriptors that can match the assessment criteria in the rating scales should be developed with the rating scale, so that the examiners have references for making more consistent judgements (Alderson et al.: 111).

In general, there are three methods of developing rating scales in language tests: intuitive, qualitative, and quantitative methods. Intuitive methods of scale development are based on principled interpretation of experience, i.e. professional judgement, whereas the other two methods involve systematic analysis of quantitative or qualitative data and are usually used in larger testing or research institutions (Luoma, 2004: 82-86). No matter which methods are used, "the basic rules for writing good skill-level descriptors […] should be brief, clear, definite and comprehensible independently" (Council of Europe, 2001: 205-207; in Luoma 2004: 82).

Marking schemes and rating scales for interpreting assessment

As late as in the first half of 1990s, there appeared to be a lack of published explications regarding assessment criteria and instruments for interpreter education (Schjoldager, 1995: 187). Interpreting assessment was done holistically by using rating scales similar to the traditional global scales in language testing.

With a formative assessment in mind, Schjoldager produced a check list style feedback sheet for assessing simultaneous interpreting "to offer an explicit, systematic

alternative to intuitive assessment procedures" where she regarded the criteria as being not only implicit but also arbitrary (1995: 194). Schjoldager's assessment criteria were divided into four categories: comprehensibility and delivery, language, coherence and plausibility, and loyalty. The students' interpreting performances were seen from the perspectives of the user, the speaker, and the listener, which is in line with the multi-dimensional perspective of the quality model for interpreting.

Interestingly, instead of fidelity and loyalty, Schjoldager's feedback sheet emphasised the importance of language and delivery skills because the interpreter's other qualities are "irrelevant" if the audience could not understand or bear to listen to the interpretation (ibid: 191). The feedback sheet did not specify, however, how much weighting was allocated to each check list item in order to produce a final mark, and the interpreting performance was still being "assessed as a whole" (ibid: 194).

Other rating scales have also been developed for interpreting assessment. They are either nominal (such as pass, fail) or ordinal (such as excellent, good, poor) scales, which cannot be averaged because each category represents some attributes of the performance under judgement (Sawyer, 2004: 105). For example, Riccardi's (2002) educational microcriteria for interpreters is a three-point ordinal scale attached to short labels like "none, some, many" or "good, satisfactory, poor", and so on. However, Riccardi's scales do not provide a mark-generating mechanism, either.

Different marking schemes were also produced by various interpreter education institutions and examination boards. For example, in Taiwan, Yang (2000) introduced a marking table for a university's professional interpreting examination, which could be used for all modes of interpreting assessments, including sight translation, consecutive interpreting and simultaneous interpreting. The main categories of the marking components, i.e. assessment criteria, were *fidelity*, *delivery*, *language*, and *time control*, each with different weightings attached, fidelity being the highest at 50%. The main

purpose of this marking table, according to Yang, was to serve as a marking tool with clear objectives for the examiners so as to enhance the fairness and objectivity of the interpreting examination (Yang, 2000: 162-163).

Yang's marking table, however, lacked rating scales and performance descriptors to match the assessment criteria. The marking table was more of a check list with different percentage weightings attached to the labels of assessment criteria, and the only quantitative component that could be measured objectively was time control. The examiners still needed to go through a deliberation process to make a consensus decision. Many interpreting examination boards still follow this practice. So an understanding of how the examiners perceive the candidates' interpreting performances and make their judgements, which is the overall research aim of the present study, will be useful to improve the procedures and design of the interpreting examinations.

Obviously, these marking schemes and feedback sheets can benefit more from systematic empirical studies. Thus, Peng (2006) produced a feedback grid as a research tool for a study on the coherence and quality of conference interpreter student training. The feedback grid was based on empirical data collected from questionnaires and interviews. The completed grid consists of four parts, or main assessment criteria: sense consistency, the language and the delivery of interpretation, and an overall judgement of the interpreter. Under each main criteria, there are sub-item criteria with a three-point-scale attached to them (Peng, 2006: 82-84). In terms of the criteria and scale format, nonetheless, Peng's feedback grid appeared to be a combination of Schjoldager's (1995) check-list feedback sheet and Riccardi's (2002) microcriteria evaluation sheet.

The various feedback and evaluation sheets for interpreting assessment reviewed above are predominantly ordinal scales similar to the numerical rating scales in language testing, with only short descriptions of the criteria in the scale. "Brevity makes scales user-friendly both for people who are reading a scale for the first time and for

assessors" (Luoma 2004: 82). However, the lack of detailed descriptors of the assessment criteria may lead to vagueness of judgements. For example, both Riccardi and Peng used simple qualifiers only to measure various features in students' interpreting performances, such as "none", "some", "many", or "mostly accurate", partially accurate", "rarely accurate", and so on. They are practical, but lack the details to help keep the subject judgement consistent.

To avoid inconsistencies in the interpretation of the short labels, it has been argued that each category or band should be furnished with descriptors of the expected performances, which refer back to the assessment criteria. Where descriptors of the rating criteria are provided, they also need to be definite. Take this criterion statement for example:

> You interpret the meaning of a sustained presentation precisely and fluently in the target language, maintaining a consistently accurate performance throughout the assignment (CILT, 2006: 25).

On the surface, this statement seems reasonable and acceptable, and is a good guideline to describing the standards for an interpreting examination. In practice, however, the description may not match what the examiners perceive when listening to simultaneous interpreting. Due to the working mode of concurrent listening and speaking, the characteristic of a conference interpreter's output is different from a normal spoken language discourse. The interpretation output is often filled with "short bursts of speech bracketed by pauses;" and "the interpreted speech often seems to be a simplified version of the original in form" (Chernov, 1979; in Liu, 2001: 8). To use the above descriptor with a rating scale for marking, therefore, it is necessary to clarify what "precisely and fluently in the target language" refer to? And to what extent a performance can be regarded as "consistently accurate?" The professional judgement clearly plays a crucial role here and the criteria need to be clearly articulated.

Peng reviewed the following set of assessment criteria statements used by an interpreter training programme in the UK's higher education system:

> To achieve 70% or higher (first class performance), a student's interpretation should:
> - show a very high degree of reliability in relaying meaning
> - be entirely coherent as discourse
> - show command of appropriate TL expression
> - achieve a standard of presentation which demonstrates mastery of the skills involved in keeping pace and addressing an audience
>
> […]
>
> To achieve 50% (the pass mark), a student's interpretation should:
> - relay meaning without systematic distortion and without major unwarranted omissions or additions
> - be mostly coherent as discourse
> - achieve a standard of TL expression which does not impede communication to a significant extent
> - achieve a standard of presentation which shows some evidence of ability to keep pace and address an audience
>
> Source: (Peng, 2006: 25)

Peng regarded these criteria statements as "implicit" and "not clearly defined"; therefore, they were not effective in separating the ability levels from the students' point of view (ibid). Many vague expressions, such as "entirely coherent", "mostly coherent", and "show some evidence", readily permitted subjective interpretation. Perhaps that was what was intended in the first place for practical reasons, but it does not help to enhance the consistency and reliability of the interpreting examinations.

Rating scales will always be subject to interpretation by the examiners (North & Schneider, 1998: 243); therefore, to help keep the groups of learners' levels apart consistently, concrete descriptions or examples are required. "Definite formulations will also support descriptor independence, so that readers will not need to read the adjacent descriptors to understand what a particular descriptor means" (Luoma, 2004: 82).

Other test instruments in SI examinations – notes and speech scripts

Perhaps in part due to the lack of proper assessment instruments, some interpreter examiners developed a convention of taking notes while listening to the interpreting performances. The interpreter examiners would then rely on the notes to make judgements or decisions. In other words, apart from some marking tables, examiners' notes might be the only important "evidence" on paper for the deliberation of the students' interpreting performances in an interpreting examination panel.

Of course, the examiners could review the recordings of examinees' performances if they are available. Recording-mediated examination marking could also reduce the stress on the examiners in a live panel examination. However, reviewing the recordings of interpretation is time-consuming. It also has attracted criticism due to the lack of *face* validity (2.3.2.b), saying that the practice does not reflect what happens in the real world because the audience listen to live interpretations and there is only one opportunity to do so in conferences (Liu et al., 2008: 19-21). No matter whether it is live marking with notes or post-examination marking with audio or video recordings, any marking procedures should aim to reduce the subjective element in the judgement process by providing definite assessment criteria for judgements.

In any case, note taking, especially by interpreters, is a very personal record of an individual's subjective perception. Different examiners may take different notes so the method is unlikely to be reliable in an examination panel. Liu et al. noted an attempt to tackle this problem, which was unique among the interpreter education institutions. One of the surveyed institutions produced verbatim speech scripts of the examination tasks to be used with a marking table. The speech scripts were segmented according to the sense units so that the examiners could rely on the scripts for more efficient and accurate note taking (Liu et al., 2008: 17), which may help reduce the memory and

cognitive load of the examiners when assessing simultaneous interpreting. The examiners' notes on the speech scripts can also be used in a more formal manner for discussing examinees' performances, such as how many errors – minor and major – each examinee had made. This means that the discussions are evidence-based and decisions could be made in a more transparent and consistent manner.

2.4.3 Subjective judgement and interpreting assessment

As mentioned earlier in 2.4.2.c, one of the methods to develop rating scales is by intuition. In fact, most language proficiency scales appeared "to have been produced pragmatically by appeal to intuition, the local pedagogic culture and those scales to which the author had access" (North & Schneider, 1998: 220). The production of the rating scales for the interpreting assessment seemed to have taken the same approach as reviewed in the previous sections. Overall, as Liu et al.'s survey study concluded, although there were common grounds, the methods and instruments for interpreting assessment varied among different interpreter teachers, examination boards and training institutions. Many rating scales and marking tables may appear to be similar, having similar assessment criteria, but their weightings are different from one marking scheme to another and their criteria statements lack definiteness and were impractical to be used in a reliable fashion (2008: 34-35).

With a pedagogical focus on peer feedback from interpreter trainees and collaboration between them, Peng (2006) also reviewed the educational standards that were used by two professional organisations – Association Internationale des Interprètes de Conférence (AIIC) and the Directorate General for Interpretation at the European Commission (formerly known as SCIC) – as well as some European interpreter educational institutions, including European Masters in Conference Interpreting (EMCI)

and Ecole Supérieure d'Interprétes et de Traducteurs (ESIT) in Paris (ibid: 22-28). Peng's observations on the assessment criteria used by those institutions were that they are "vague", "required to restructure the benchmarks systematically", and "leave much room for subjective judgement to form" (ibid: 24). She found that the School of Translation and Interpretation (ETI) in Geneva has a list of assessment criteria that is "more structured and more explicit"; however, "scope for confusion" still exists (ibid: 25). These observations were not too far from those in Liu et al.'s (2008) survey of the eleven interpreter schools as discussed earlier.

Liu et al.'s survey study reveals a serious concern about the lack of consensus within the interpreter education community on the assessment methods for interpreting; each does things in their own way. Although there have been efforts to improve the assessment methods, the approaches have often been based on subjective judgement rather than on empirical data (2008: 17-18). As mentioned earlier in 2.4.1, one benefit of producing the test specification is that when writing the document test designers have opportunities to discuss and exchange opinions. Thus, different viewpoints regarding the purpose, nature as well as the format of the test may be reconciled in the process. Therefore, the consensus building in the endeavours to develop better interpreting examinations should begin from the production of the test specifications.

* * *

This section has reviewed some of the lessons from the language testing discipline, such as the production of test specifications, the design of the rating scales as well as the studies that improve understanding of how the examiners judge and use the scales. They are particularly useful for the development of interpreting assessment as it tries to ground itself on explicit, empirically supported criteria and methods for testing. The approach to study these issues in the field of the interpreting assessment and the methodological implications will be considered next.

2.5 Methodological implications to the interpreting assessment

This section considers the methodological implications to the interpreting assessment by reviewing some lessons learned from language testing (2.5.1), and presents the multi-strategy research approach for the present study (2.5.2 and 2.5.3).

2.5.1 Rater effect: lessons from language testing

By using some psychometric approaches, such as Generalizability Theory and many-facet Rasch measurement [10], researchers in various performance settings statistically modelled and demonstrated "the pervasive and often subtle ways in which raters exert influence on ratings" (see e.g. in Eckes, 2005: 198). These subtle ways of influences are referred to as the *rater characteristics* or *rater effect*.

Rater characteristics were conceptualised "in terms of the difference between an idealized judge (the 'perfect' examiner) and actual judges ('ordinary' examiners)" (Lumley & McNamara, 1993: 3). A perfect examiner that is always consistent and reliable is almost impossible to find, and it is the ordinary examiners that have problems, such as halo, overall severity/leniency, central tendency, and random errors in their judgement (ibid: 3). These problems, i.e. rater effect, will have an influence on the results of many performance-based assessment, making the assessment procedure become unreliable and threatening the validity of the test (Eckes, 2005: 197).

Halo effect occurs when raters are unduly influenced by a single aspect in a test performance and inappropriately generalise it to all other aspects in the performance.

[10] Generalizability Theory, or G Theory, is a statistical theory for evaluating the dependability ("reliability") of behavioural measurements (Shavelson, 2004). As in G Theory, multifaceted Rashch analysis lets the researcher look at a range of facets and how they contribute to score variance (Fulcher, 2003: 210-213).

This effect tends to reduce raters' evaluations "to a single overall impression rather than specific areas of competence. Such lack of differentiation resulted in very little variation within the ratings given in distinct categories" (Phelps, Schmitz, & Boatright, 1986: 151). Rater severity or leniency happens when examiners exhibit an "inability to discriminate between differing degrees of quality performance" (ibid: 152) and consistently score students at the two ends of the rating scale, whereas with central tendency, examiners tend to consistently rate students at the midpoints of a scale.

Being a performance-based assessment, examiners in language testing, and interpreting assessment in the case of the present study, are not immune to the rater effect. As "the reliability of any test of spoken language hinges on the role of oral examiners or raters" (Breeze, 2004: 2), many empirical studies have been carried out to understand the effect of the role of examiners in language testing (Bachman et al., 1995; Eckes, 2005; Fulcher, 2003; Lumley & McNamara, 1993; Upshur & Turner, 1999) so that "human errors", i.e. the unsystematic test errors, can be reduced by applying suitable examination procedures, such as the training of examiners that allows the examiners to become familiar with the marking systems and apply them consistently (Alderson et al., 1995: 105).

The training of examiners can "reduce the random error in rater judgements" and make the examiners more self-consistent (Lumley & McNamara, 1993: 3). However, research studies found that rater severity or leniency would still exist even after specific rater training (Eckes, 2005: 198-199). Therefore, researchers in language testing hold the view that the function of the training of examiners is not "to force raters into agreement with each other (interrater reliability), but rather to train raters to be self-consistent (intrarater reliability)" so that "it allows for some variability in rater reactions" to the test performances (Weigle, 1998: 265), i.e. the examiners can have some room to assess in a natural way. That is to say, rater training can reduce "extreme

differences" in rating behaviours, but the rater variability cannot be eliminated; therefore, "compensation for rater characteristics needs to be built into the rating process" (Lumley & McNamara, 1993: 3). In order to do so, test designers need to identify "sub-patterns in the behaviour of raters which may be systematic in some way, that is, predictable, and thus able to be compensated for" (ibid).

The approach that the researchers in language testing took to identify patterns of rater behaviours was to observe how the examiners, or raters, interact with various facets of the examinations, such as *rater–ratee* interaction, *rater–task type* interaction, and *rater–criteria* interaction (Eckes, 2005: 199). For example, "raters may display particular patterns of harshness or leniency in relation to only one group of candidates, not others, or in relation to particular tasks, not others, or on one rating occasion, not the next" (Lumley & McNamara, 1993: 3), and so on. In language testing, these interactions were studied mainly by using the statistical method of multi-faceted measurement, such as the aforementioned generalizability theory and many-facet Rasch measurement. See Fulcher (2003: 210-215) for a detailed introduction of these methods.

The lessons in language testing may give useful groundwork for considering the methodological approach to carry out research studies on similar issues of the interpreting assessment, such as the present study.

2.5.2 Seeking an approach for studying the interpreting assessment

As reviewed in this chapter, research in the field of the interpreting assessment is still at the initial stage of exploration. Based on the experiences in the discipline of language testing, understanding how examiners make judgement and developing effective test instruments based on valid test constructs are essential ground work for both practical examination administration and research studies on the issues

surrounding the interpreting assessment. However, some difficulties will arise in the interpreting assessment if psychometric methods are used to study examiner-related issues. To begin with, for example, in order to make statistical modelling of rater characteristics possible, the examiners need to be trained so they are firstly internally consistent when assessing examinees by using a suitable test instrument, such as a rating scale (McNamara, 1996: 127; in Eckes, 2005: 199). Using a flawed rating scale in a study that employs psychometric method will impose higher limitations in generalising the research findings (Caban, 2003: 34).

Furthermore, in the case of the interpreting assessment as discussed in 2.4.3, there is still little consensus on a standard assessment procedure and assessment instrument for the interpreting examinations. It would not be ideal to base a research study on a rating scale and an examiner training session of the interpreting assessment 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: 5), and encountered some problems that could not be addressed solely by using the quantitative-oriented psychometric research method (Upshur & Turner, 1999: 103-107). Qualitative research approach was suggested to supplement the statistical method. For example, protocol and interview analysis was proposed to complement rater performance reports *before* the FACETS[11] analysis "to see if it is possible to identify beforehand the likelihood of personal circumstances influencing a rater's severity or leniency on a particular occasion" (Lumley & McNamara, 1993: 16).

Brown conducted a study on rater variables in a language speaking test, and found that "there were no significant differences between the different types of rater in terms

[11] FACETS is a computer software for many-facet Rasch measurement, a statistical method.

of the overall grade awarded. However, there were significant differences in ratings awarded for some individual criteria" (Brown, 1995: 1). In other words, the examiners have "different perceptions of what constitutes good performances" (ibid). These perceptions in the mind of the examiners can only be identified and understood by using qualitative methods, such as an interview analysis. Therefore, as Caban suggested, "Raters could be interviewed and a qualitative analysis might be performed to determine categories of importance", i.e. different weightings of assessment criteria on the rating scale (Caban, 2003).

In studying tests of second language speaking, Fulcher pointed out that "there is almost always a qualitative element present, especially when making judgements about the meaning of statistics" (2003: 216), and "Quality approaches provide insights that cannot be gained from statistical analysis. The methods […] provide insights into how experts make judgements (critical at all stages of test development)" (ibid: 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.

Bryman used the term *multi-strategy research* to refer to "research that integrates quantitative and qualitative research within a single project" (Bryman, 2004: 452). He suggested that multi-strategy research "may provide a better understanding of a phenomenon than if just one method had been used," and "must be competently designed and conducted" (ibid: 464). Pollitt and Murray successfully demonstrated the usefulness of a study design that combined qualitative and quantitative research methods to elicit and validate the constructs of the rating scale for speaking test (Pollitt & Murray, 1996: 88). They employed Thurstone's Method of Paired Comparisons, a quantitative approach, "to monitor consistency within individual raters and between raters" (ibid: 79), which also "facilitated the expression by the judges of the aspects that seemed salient to them" (ibid: 88) for analysis that was focused on quality. The

multi-strategy research approach adopted by Pollitt and Murray 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 record and analyse the qualitative data systematically.

2.5.3 Rationale of research method

For the purpose of the present study, which studies the examiner-related issues in the interpreting assessment, one most useful aspect of the Paired Comparison method is that it does not require a rating scale. Therefore, inspired by Pollitt and Murray's study, to meet the research aims and objectives stated in Chapter 1, the design of this research study takes a multi-strategy approach, or a mixed-method approach, by employing quantitative methods, i.e. the Method of Paired Comparisons and some common statistical methods, to establish a framework for analysing qualitative data, i.e. interview comments, that is generated alongside the paired comparison method. Apart from the quantitative analytical framework, the interview comments will also be coded and analysed by basing on the Grounded Theory (GT) which is a qualitative research methodology; however, only part of the GT process, i.e. coding of interview comments, is adopted because the interview data will be collected only once in this study.

The theoretical aspects of the above identified research methods are reviewed below before presenting the overall research design and methodology of this study in Chapter 3.

2.5.3.a Thurstone's Method of Paired Comparisons

This section reviews the fundamentals of the Method of Paired Comparisons and shows how it is appropriate and useful for the purpose of this study.

Scientists use all kinds of physical scales to measure objects, such as weight, height and length. For qualitative judgement on human behaviour, however, the measurement is often more subjective and requires a different approach to achieve a more objective measurement. In psychology, *indirect* scales are used by psychologists to measure people's mind and attitudes, such as rating scales questionnaires or interview techniques, and psychological scaling models can be distinguished according to whether they are intended to scale persons, stimuli, or both.

The Likert scale, for instance, is a *person*-centred approach because the scale scores are mainly used to indicate the attitude of the person completing the scale to the stimuli on the scale. By contrast, Thurstone scaling is a method used to measure the *stimuli* evaluated by a group of people with respect to some designated attributes. In other words, it is the stimuli rather than the persons that are scaled by Thurstone's scaling method (McIver 1981, Togerson 1958 in Li, Cheng, Wang, Hiltz, & Turoff, 2001). In this study, the stimuli would be the student interpreters' performances judged by the examiners. I intended to use the measurements of the stimuli from the Thurstone scaling method to monitor the consistency of the examiners' judgements.

Thurstone's own experiment may be used to illustrate the point regarding the measurement of the stimuli. He selected 19 different criminal offences as the stimuli and asked 266 subjects to compare the seriousness of these offences in pairs. Then, based on the comparisons and by applying a set of mathematical equations, he produced a "Scale of Seriousness of Offences" that ranked the 19 offences with the most serious ones like rape and homicide on the top to the less serious offences like receiving stolen goods and vagrancy at the bottom of the scale. Thurstone's scaling technique was able to show the composite judgement (or attitude) of the group of 266 people with credible measurement of the relative "distances" between each offence on the scale (Thurstone, 1959: 67-81). The distance between rape and homicide was different, for example, from

the distance between receiving stolen goods and vagrancy on the scale. It is like reading a map with many towns and cities positioned on a motorway route; the distances between each town and city are all different.

If the Likert scale had been used in the experiment, the subjects would have been asked to rank the seriousness of the 19 offences, i.e. the stimuli, in *equal* distance on a scale of 1 to 5, for instance, in stead of comparing them in pairs. The results would have been how many percentages of people regarded offence A as more serious than the other offences on a 1-to-5 scale, or the offence was regarded as serious at certain level on the scale by the group of subjects, and so on. In other words, it is the subjects that were being scaled in relation to the 19 offences. The results of the Likert scale may be more like a demographic map that show different populations of different towns and cities. It would be difficult for the Likert scale to show the *relative* distances with credible measurement between each of the criminal offences in the minds of that specific group of people as a whole (Li et al., 2001).

As each stimulus is paired with every other one to compare, Thurstone (1959) referred to his scaling technique as the Method of Paired Comparison, and said that his intention was "to apply the ideas of psychophysical measurement in the field of social values" (Thurstone, 1959: 67). His scaling method did not assume that each stimulus always evokes the same discriminatory process for different individuals or even for the same individual at different times (Li et al., 2001). The attributes that form the focus of the discriminatory process, i.e. an individual's judgements, may be aesthetic, ethical, or linguistic in nature (Pollitt & Murray, 1996: 78). Therefore, the scaling technique is able to transform rank order data or comparative preferences by a group of individuals into a single composite interval scale, such as the Scale of Seriousness of Offences.

The basis for the paired comparison method dates back to its first reported use in the mid-1800s. Although the technique is a very powerful approach for producing a

highly reliable ranking of the rated items, it is underutilized by survey researchers due to the amount of data that often must be gathered, and thus its cost and the burden it places on respondents[12]. Despite its underutilization by survey researchers, nevertheless, the paired comparison method is widely used in commercial sectors to prioritize a range of options or root causes (those vital and systemic for improvement) for its reliability when the number of items (i.e. people or options) to be scaled is small, such as in a small-and-medium-sized company.

Considering the size of this present study, therefore, and the fact that the scaling technique does not need to use a separate rating scale for making the paired comparisons, Thurston's Paired Comparison Method is a useful research tool for this study. Examiners' judgements on the students' interpreting performances, i.e. the stimuli, can be compounded into a single scale of interpreting skill proficiency, showing the judged ability levels of students' performances in measured positions on the scale. Therefore, the consistency levels of the examiners' judgement can be monitored by comparing the results of the compounded scales that represent the examiners' judgements on the same student interpreters' performances, for example, between different groups of examiners or between different assessment occasions. For this reason the Method of Paired Comparisons was used in this study for data collection at the start of the process of the examination simulation (see 3.1 and 3.2).

2.5.3.b Interview comment of subject examiners

As the research questions also look for the reasons that lie behind the examiners' judgements, the subject examiners would be interviewed during the paired comparison process so their comments could serve as a window that enables us to look into the

[12] Reference source: SAGE Research Methods Online, http://srmo.sagepub.com/view/encyclopedia-of-survey-research-methods/n362.xml). Accessed 25 Jan. 2011.

examiners' perceptions and applications of the assessment criteria.

Ideally for the purpose of this study, it would be best to ask the examiners to give a concurrent introspective verbal report while listening to the students' interpretations. However, due to the nature of the simultaneous interpreting examination process, such a procedure was not practical. It is unlikely that examiners could "think aloud" the judgement process while listening to the simultaneous interpreting.

In addition, for validity reasons, the design of this study needed to allow the subject examiners to work as normally as possible in a simultaneous interpreting examination while at the same time allow the researcher to capture the examiners' thoughts during the assessment process. Therefore, the subject examiners would simply be asked to comment on the students' interpreting performances immediately after listening to them and when comparing them in pairs.

The subject examiners were interviewed by following a guideline (see 3.2.4) so the interviewing process could be regarded as semi-structured. The process is flexible, allowing the interviewees, i.e. the examiners, to make comments without much interference, so that what the examiners viewed as important in explaining and understanding the issues involved in assessing the student interpreters could be explored (Bryman, 2004: 321). Therefore, examiners' comments during the judgement process would be recorded as research data for analysis.

2.5.3.c Grounded Theory and its coding practice

Grounded theory methods emerged from sociologists Glaser and Strauss in the 1960s, who "aimed to move qualitative inquiry beyond descriptive studies into the realm of explanatory theoretical frameworks" (Charmaz, 2006: 4). Essentially, Grounded Theory advocates that theories should be *developed* from research studies that are grounded in data instead of *deducing* testable hypotheses from existing theories.

Qualitative data, such as interviews, are subjected to systematic analysis to identify abstract and conceptual understandings of the studied phenomena. The qualitative data analysis follows certain guidelines to extract concepts that can be further compared and categorised for analysis and generating theories. Therefore, the theories from such study approach maintain a strong foundation in data, i.e. *grounded* in data (Charmaz, 2006: 4-6; Glaser & Strauss, 1967).

Researchers may adopt and adapt basic grounded theory guidelines, or tools, such as theoretical sampling, coding, memo-writing, theoretical saturation and constant comparison, for conducting a diverse range of studies (Bryman, 2004: 401-403; Charmaz, 2006: 9). Charmaz regarded grounded theory methods as "a set of principles and practices, not as prescriptions or packages", which can "complement other approaches to qualitative data analysis, rather than stand in opposition to them" (ibid: 9). Ground Theory, therefore, may also work well with multi-strategy, or mixed method, research study designs, such as the current study.

For example, the interview comments in this study, i.e. the qualitative data, may be transcribed for analysis. In order to scrutinise the unstructured transcribed data, a systematic approach, i.e. Grounded Theory's coding practice, is adopted to generate theoretical concepts. There are three distinctive types of coding practices in the Grounded Theory – open, axial, and selective coding (Bryman, 2004: 401). Open coding involves line-by-line examination of the text data, breaking them down, and comparing them, and from these coding processes emerge *concepts*; similar concepts then can be further grouped into *categories*. Axial coding makes connections between categories to identify any pattern, whereas in selective coding, core categories are selected and related to one another to explain their relationships. In this study, the emerged concepts and categories from the qualitative data will be the assessment criteria and/or the test constructs of interpreting examinations.

The Grounded Theory approach is an iterative process for collecting and analysing qualitative data (such as interviews) until a theory is derived from or "grounded" in the data (ibid: 401). The current study, however, did not go through the iterative data collection process because the study design only allowed one data collection event, i.e. the paired comparison comments. Nevertheless, the coding principle of recursively analysing the qualitative data – a key process in the Grounded Theory approach – remains useful for the systematic analysis of verbal comments. The qualitative results may then be triangulated with the results of quantitative data to explore the relations between the assessment criteria and the examiners' judgement results.

Next in Chapter 3, the overall study design and methodology of this research project is explained.

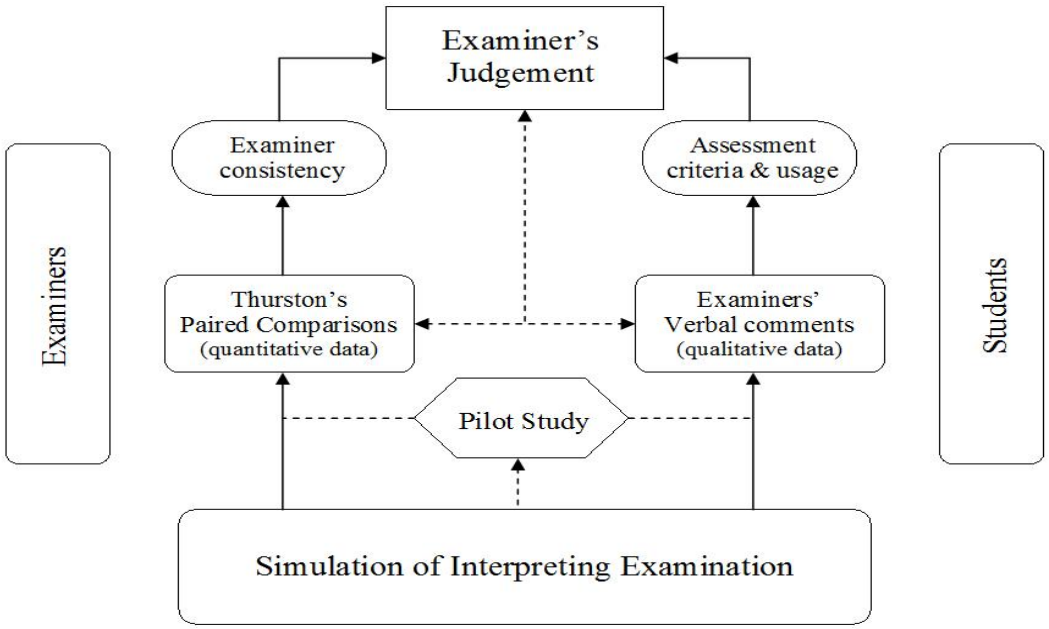
CHAPTER 3

A Multi-Strategy Approach

3.0 Design of the research study

A simulated examination of simultaneous interpreting was conducted to collect data for answering the research questions, and to achieve the aims and objectives stated in Chapter 1 (1.3). Figure 3-1 illustrates the design of the present research study.

Figure 3-1 Research study design



Two research methods were identified for data collection: Thurstone's Method of Paired Comparisons monitored judgement consistency through the examiners' rankings of the student interpreters; examiners' verbal comments during the paired comparisons were recorded as data for qualitative analysis in relation to the student ranking results. These quantitative (student rankings) and qualitative (examiners' comments) data collected from the examination simulation were to be recorded, analysed and cross-checked to reveal the decision making process of the examiners when assessing simultaneous interpreting, and to answer the research questions of the study.

3.1 Pilot study

A small-scale study was conducted for pilot-testing the usefulness of the above chosen research methods. In the pilot study, I would like to ascertain whether the methods could (1) successfully monitor the consistency of examiners' judgements (Method of Paired Comparisons), and (2) effectively elicit examiners' criteria when assessing simultaneous interpreting performances (interview comments). The pilot study was also (3) a test run of the overall procedures of the examination simulation for the data collection, on which the main study design would be based.

Detailed report of the pilot study was presented in an article of a book (Wu, 2010). This section shall summarise the procedures of the pilot study and its main findings for improving the main study design.

3.1.1 Study procedures of the pilot study

The pilot study took the format of a simulated examination of simultaneous interpreting. Eight examiners – five interpreter teachers and three English language and translation teachers – were invited to participate as subjects in the pilot study. The selection of different profiles of participant examiners was to compare interpreter and non-interpreter examiners' judgement patterns to see if there were differences. As the main subject of this study is on interpreters, more interpreter examiners were recruited. The examiners were coded P1 to P8. All of them were female native speakers of Mandarin Chinese.

The examiners were asked to assess and compare five postgraduate student interpreters by viewing and listening to their authentic examination video recordings,

which were selected according to the students' marks in the examination in their Year-1 study in a two-year MA in Interpreting study programme in a UK university. The students were coded as S1 to S5; all of them were females.

Three students had Chinese as their A Language and English as B Language[13]; one student had a language combination of Chinese B and English A, and one had Chinese C and English B. The reason to include two non-Chinese A students was in the hope to illicit a wider range of assessment criteria from the examiners.

The interpreting task was from English into Chinese. To limit the time constraints and efforts of the participant examiners, only the first three minutes of the speech task were used. If the task was too short, it might not be able to generate enough data. If a longer task was used, however, there might be risks of data overload; the data collection process might also be too long to keep the participant examiners' concentration in a good condition. Therefore, the pilot study also aimed to verify the suitability of the length of the task for the main study. (see 3.1.2 below).

There was only one examiner in each session of the examination simulation. 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 in the study, there were ten pairs to compare. The comparison rota for students S1, S2, S3, S4, S5, for example, was as follows: after viewing performances of S1 and S2, examiners were asked to compare these two. Then the examiner would view S3 and compare S3 with S2. After comparing S3 with S2, the examiner would view S4, and then compare S4 with S3, then view S5 and compare with S4, then compare S5 with S1, S1 with S3, and so on. This process would continue until all ten pairs were compared.

[13] A Language refers to an interpreter's most active language, usually the mother tongue. B Language is an interpreter's second active working language, usually an acquired second language at near-native level. C Language usually is an interpreter's second foreign language. An interpreter may use his/her C Language to receive information, but may not be able to use it to convey messages with confidence. See AIIC web site at www.aiic.net (accessed 12 August 2009).

Immediately after viewing each pair, the examiners were asked to decide which was better – a comparative Thurstonian decision – but not to give marks to each performance; then the examiners were asked to comment on the performances, in what way they were better or worse, similar or different and any other relevant comment. The comments were recorded digitally into the computer[14] for later analysis. After comparing the ten pairs, the examiners were then asked to give their overall rankings of the five student performances.

Finally, the participant examiners were invited to give comments on the appropriateness of the setup of the examination simulation and the overall study procedures, including the length of the examination task, which would be used to improve the main study design.

3.1.2 Data analysis and findings of the pilot study

The data collected from the pilot study were (a) the results of the paired comparisons and the overall judgement rankings, (b) the audio recordings of the examiners' comments, and (c) the participant examiners' comments on the study procedures. Data (a) were analysed to monitor the consistency of the examiners' judgements and to ascertain the usefulness of the research method. The audio comments on student performances, i.e. (b), were transcribed into text and analysed to see if assessment criteria used by the examiners could be extracted. Finally, the comments on the procedures (c) were considered to refine the design of the main study.

3.1.2.a Results of the paired comparisons and overall judgements

The results of the paired comparisons (hereafter PC) are presented in Table 3-1,

[14] A PC microphone and software (Audio Cleaning Lab, http://www.magix.com) were used to record.

which shows the PC winners of the eight examiners' choices. As the main study would also use the same approach to analyse the PC data, the analysis procedures are reported below in detail so the process of data treatment is clear.

Table 3-1 Winners of paired comparisons

Examiners	Student pairs & Comparison winners									
	S1-S2	**S2-S3**	**S3-S4**	**S4-S5**	**S5-S1**	**S1-S3**	**S3-S5**	**S5-S2**	**S2-S4**	**S4-S1**
P1	S2	S3	S3	S4	S5	S3	S3	S2	S2	S4
P2	S1	S3	S4	S4	S1	S3	S3	S5	S4	S4
P3	S2	S3	S3	S4	S5	S3	S3	S2	S4	S4
P4	S2	S3	S3	S4	S1	S3	S3	S2	S4	S4
P5	S2	S3	S3	S4	S1	S3	S3	S2	S2	S1
P6	S2	S3	S3	S4	S1	S3	S3	S2	S2	S4
P7	S2	S3	S3	S5	S5	S3	S3	S5	S2	S4
P8	S2	S3	S3	S5	S5	S3	S3	S5	S2	S1

P1-8: examiners, **S1-5**: students

Analysis procedures of the paired-comparison data

Table 3-2.a Paired comparison rankings

PC Ranking points	S1	S2	S3	S4	S5
P1	1	4	5	3	2
P2	3	1	4	5	2
P3	1	3	5	4	2
P4	2	3	5	4	1
P5	3	4	5	2	1
P6	2	4	5	3	1
P7	1	3	5	2	4
P8	2	3	5	1	4

Ranking point: 5 – 1st place, 4 – 2nd place, 3 – 3rd place, 2 – 4th place, 1 – 5th place

Table 3-2.b Overall judgement rankings

OJ Ranking point	S1	S2	S3	S4	S5
P1	1	4	5	3	2
P2	4	1	3	5	2
P3	1	3	5	4	2
P4	2	4	5	3	1
P5	3	5	4	2	1
P6	2	4	5	3	1
P7	1	3	5	2	4
P8	4	2	5	1	3

Ranking point: 5 – 1st place, 4 – 2nd place, 3 – 3rd place, 2 – 4th place, 1 – 5th place

In the ten comparisons made per examiner, the maximum number of times a student could win is four, and the minimum is zero. Because zero is not computation-friendly, for ranking conversion and statistical calculation, the 1st place winner (who won four times per examiner) is assigned a ranking point of 5; the 2nd place (won three times) a ranking point of 4; 3rd place (won twice), 3; the 4th place (won once), 2; the 5th place (no win), 1. There was no tied comparison. The converted PC ranking points are shown in Table 3-2.a. As for the overall judgement (hereafter OJ) rankings, no conversion was needed because each examiner had given a direct ranking of the five students' performances. The OJ rankings are shown in Table 3-2.b. These rankings were then used for statistical computation to produce the Thurstone's scales.

Validating the Thurstone scales for this research study

The first objective of the pilot study was to verify the usefulness of the Thurstone's Method of Paired Comparisons. This section first presents two Thurstone scales, and then examines their reliability by using Cronbach's alpha, a statistical correlation coefficient for estimating scale reliability. The usefulness of the scales will be validated by referring to two sources of external evidence: the examiners' interview comments and the original marks given to the five students in the actual examinations.

Figure 3-2.a Thurstone scale – paired comparisons (PC)

← Better	Scale of Interpreting Competency			Worse →
S3	S2	S4	S5	S1
1.19	0.08	0	-0.55	-0.71

Figure 3-2.b Thurstone scale – overall judgement (OJ)

← Better	Scale of Interpreting Competency			Worse →
S3	S2	S4	S1	S5
1.03	0.16	-0.08	-0.47	-0.63

Two Thurstone scales (Figures 3-2.a and 3-2.b) were produced based on the ranking points in Table 3-2.a and Table 3-2.b respectively. As explained in 2.5.3.a, Thurstone's scaling technique transforms rank order data and comparative preferences by a group of individuals into a single composite interval scale. In the case of the study here, the scales indicate the relative positions of the five students on the two occasions (PC and OJ) based on the aggregated judgements of the eight examiners. The two Thurstone scales (hereafter T scales), therefore, can be regarded as the five students' interpreting proficiency scales; positions further to the left indicate better performances, further to the right, worse performances.

The numbers under the student codes are the scale values. Using the scale values, the two T scales were compared by computing the Cronbach's alpha. The Cronbach's alpha for the two T scales is 0.987. The result suggests that the consistency level of the two T scales could be considered as excellent[15]. In other words, the examiners, as a group of eight, were judging the five students at a consistency level that is statistically excellent between the two occasions. On the two T scales, the ranking orders of the five students' interpreting competency, from better to worse, are basically the same with S3 shown as the most competent, followed by S2 and S4, S5 and S1. Although the order of S5 and S1 is reversed on the OJ T scale, the difference in the scale values is small.

Two sources of external evidence – the examiners comments and the original examination marks that the students received – could be used to verify the validity of the T scales in monitoring the examiners' judgements. On the two T scales, for example it is worth noting that S2 and S4 are in a closely ranked pair as are S1 and S5, while S3 is ranked markedly higher than the two pairs. These positions on the T scales are echoed by examiners' comments, indicating that it was difficult when comparing S2 with S4

[15] George and Mallery (2003) provide the following rules of thumb regarding Cronbach's alpha reliability coefficient: "_ > .9 – Excellent, _ > .8 – Good, _ > .7 – Acceptable, _ > .6 – Questionable, _ > .5 – Poor, and_ < .5 – Unacceptable" (George & Mallery, 2003: 231).

and S1 with S5. This might be due to the similarity in levels of performances so the students appear in close pairs. In contrast, most examiners commented that S3 was far better than the others and so appears at a markedly leading position on the scales.

The actual examination marks of the five students (in parentheses below) also supported the validity of the T scales. The marks matched the students' positions well on the T scales with S3 (73) on the lead far ahead of the others, followed by S4 (65), then with S1 (63) and S5 (62) as a close pair. S2 (56) was the only one out of place in terms of the mark vs. T scale positions. This was not unexpected because the marks were given by a different panel of examiners in the actual examination. It is the objective of this research study to identify the reasons behind the inconsistency in the examiners' judgements such as this.

All in all, supported by the examiners' comments and the references to students' original marks of the examination, the results above indicate that Thurstone's Method of Paired Comparisons and scaling technique is a useful tool to monitor the examiners' judgements on the five students in this study.

Examiner consistency problem

A cursory analysis of Tables 3-2.a and 3-2.b, however, reveals some problems: *individually*, the eight examiners' judgments of some the students fluctuated widely. For example, some students (S2 and S4) were judged to be the best as well as the worst performers at the same time by different examiners. It was also found that the non-interpreter examiners achieved a better consistency level in their judgements compared to the interpreter examiners (Wu, 2010). Due to the small sample size in the pilot study, these results may not be generalised. However, they gave evidence to support the hypothesis of this study (1.3), i.e., the examiners' judgements may fluctuate to a degree that causes concern in a simultaneous interpreting examination.

In their study on developing the rating scale for speaking tests, Pollitt and Murray reported that non-language specialist examiners demonstrated an impressive level of consistency in assessing language students' oral proficiency, and suggested to repeat the study with language specialist examiners to see if they could match the consistency level of the non-language specialist examiners (Pollitt & Murray, 1996: 88). In the case of the present research study, therefore, it would be useful to also include non-interpreter examiners, whose consistency level was observed as excellent in the pilot study, as a contrast group in the main study to show reasons for the different consistency levels between different groups of examiners.

3.1.2.b Results of the examiners' comments

Subject examiners' comments during the comparison process were recorded and transcribed for analysis. The commentary data were assessed to see if it could (1) yield the assessment criteria that the examiners used during the examination simulation, and (2) identify possible reasons why the examiners judged differently or similarly.

Extraction of assessment criteria

Verbatim transcripts of all eight examiners' comment recordings were made and examined line by line. Based on the coding practice of Grounded Theory (2.5.3.c), key words/phrases or concepts in the transcriptions were extracted and coded. Then, similar key concepts were further sorted into a total of five groups (also see 3.2.5 and 5.1), which were labelled according to the common nature of the concepts in each group, such as Fidelity, Completeness, Delivery, Strategies, and Others which contained concepts that did not fit well in the other four groups (Wu, 2010).

These groups of concepts represented the main considerations that the examiners showed when comparing student interpreters' performances, and were also broadly in

line with the quality standards reviewed in Chapter 2. Therefore, it is reasonable to regard them as the assessment criteria used by the examiners, and hence the usefulness of the research method of the interview comments. These five groups of key concepts as assessment criteria would also be useful as a basic framework or guidance for data analysis in the main study.

Inferring the reasons for the inconsistencies

In addition to extracting the assessment criteria, a wealth of information on how the examiners reached their decisions was also revealed in their comments, which were mostly in the group labelled as "Others". Examining their comments during the paired comparisons, could allow the identification of common criteria as well as discrepancies in their using of the criteria. This would be useful to answering the research questions. However, given the amount and the nature of unstructured qualitative data of the examiners' verbal comments, the qualitative data were not fully explored in the pilot study; only three examiners were selected to compare their summary comments and decisions for analysis (Wu, 2010: 318-323). A more systematic approach would be required to process the information in the main study with an even larger number of subject examiners, which will be explained later in 3.2.

3.1.2.c Feedbacks on procedures and stimulant materials

The use of video recordings to assess student interpreters' performances was considered adequate by all eight participant examiners. Four examiners expressed concerns about the "fairness" of including a student with a C Language in the examination. Therefore, in the main study the C-Language student was replaced with a new student whose A Language is Chinese, reducing the gaps of differences among the five students so this study may explore finer comparisons among the judgements of the

participant examiners (also see 3.3.2). Another student's Chinese was her B Language; most examiners, except one, did not suspect that Mandarin Chinese was not her A Language, so the Chinese B student was kept for the main study.

The time for each study session ranged between 1.5 and 2.5 hours. Five participant examiners said that their concentration level would deteriorate if the process exceeded two hours. Seven examiners were happy with the length of the examination recording. One interpreter examiner commented that using a 15-minute long task could evaluate certain interpreting abilities, such as the sustaining ability – or the "gritting power" as she put it – to manage a longer period of simultaneous interpreting, but she had no objection to using a shorter examination task. Considering the time constraint and the examiners' effort required, the researcher decided that a three-minute examination task was sufficient for the purpose of this research study.

The procedures for the examination simulation and paired-comparison comment in general were considered clear and adequate by all participant examiners. The combined method of paired comparison with interview comments was found to be useful for its flexibility, allowing very little interruption as the examiners made comments when making their comparative decisions. A more detailed procedural briefing session was planned based on the experience of the pilot study.

It was also noticed that the data recording method could be further improved, such as recording the winners of the paired comparisons and the examiners' overall rankings of the five students. To cope with extra examiners in the main study and to ensure accurate data entry, a more systematic method was necessary. Therefore, worksheets were designed after the pilot study to remove this potential problem (see 3.2.3 Data Recording and Storage).

Finally, the pilot study highlighted that non-interpreter examiners preferred to use the speech script when evaluating, whereas interpreter examiners preferred just listening

and taking notes on a blank sheet of paper as reviewed in Chapter 2. Therefore, the script use during the interpreting examination would be noted in the main study to see if or how it affects the examiners' judgements.

3.2 Main study

With the findings in the pilot study, the main study procedures were fine-tuned and improved. It was carried out based on the research study design in 3.0. This section presents the characteristics of the participant examiners (3.2.1), student examination recordings as stimuli for study (3.2.2), data recordings and storage (3.2.3), procedures of the main study (3.2.4), and the strategies for data analysis (3.2.5).

3.2.1 Participant examiners

As the present study does not intend to make direct inferences about the general population of interpreting examiners, a non-probability sampling technique – purposive snowball sampling – was adopted to recruit participant examiners for this study. The researcher initially made contact with colleagues in the profession, i.e. interpreters, interpreting and/or translation teachers, and these respondents further identified others who belong to the target population of interest and subsequent respondents were selected based on the referrals.

To balance out the possibility that participant examiners would recommend like-minded peers, the "seeds" for snow-balling were initially selected in three different backgrounds: non-interpreter examiners, interpreter practitioners in the market, and interpreters who also teach simultaneous interpreting in universities and colleges. As the focus of this study is on the interpreter examiners, the number of non-interpreter examiners was limited to about half the number of interpreter examiners.

In total, thirty examiners were recruited for the main study, including five who had participated in the pilot study. Table 3-3 shows the characteristics of the participant

examiners of this study. Among the thirty examiners, twenty-four were female and six were male. Twenty-eight of the subject examiners were based in Taiwan and two in the UK at the time when the study was conducted[16]. The working languages of the subject

Table 3-3 Characteristics of participant examiners in main study

Examiner code name	Interpreter: 🎧	Translator: 🖋	Teaching SI: 🗣	Using script†: 📄
R1		🖋		📄
R2		🖋		
R3		🖋		
R4	🎧		🗣	
R5	🎧		🗣	
R6	🎧		🗣	
R7		🖋		📄
R8	🎧		🗣	📄
R9	🎧			
R10	🎧			
R11		🖋		📄
R12	🎧		🗣	📄
R13	🎧			📄
R14	🎧		🗣	
R15		🖋		📄
R16	🎧		🗣	📄
R17*	🎧		🗣	📄
R18	🎧		🗣	📄
R19	🎧			📄
R20	🎧		🗣	
R21*	🎧		🗣	📄
R22	🎧		🗣	📄
R23	🎧		🗣	📄
R24	🎧			📄
R25*		🖋		📄
R26*		🖋		📄
R27		🖋		📄
R28*	🎧			📄
R29		🖋		📄
R30		🖋		📄
Total:	**19**	**11**	**13**	**21**

*Subject examiners who also participated in the pilot study. †The script usage reported in the table was compiled from the research field notes after the examination simulation.

[16] Preliminary scrutiny of the data indicates that there is no significant difference in the judgement patterns between male and female examiners, or between the geographical bases of the examiners. Therefore, the gender and geographical bases are not indicated in Table 3-3 for confidentiality considerations of the subject examiners' identity.

examiners are all the same – Mandarin Chinese and English. Mandarin is the first language of all the examiners except one who was based in the UK; the language attribute of examiners is not presented here for identity confidentiality consideration.

The 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

"Professional" here means that at the time when the study was conducted the person had been working in the field of interpreting, translating or both fields for over three years. "Substantial teaching experience" here means that the person had had at least three years of full-time teaching experience in a university or college. As noted in the pilot study there was a difference in the preference of using speech script in the interpreting examination. As the use of assessment instrument might affect the way examiners assess students' interpreting, this feature is also included in Table 3-3 for analysis.

The examiners were informed in general terms what the study was designed to achieve, such as to understand how the examiners assess the student interpreters, when they accepted the invitation to participate in the examination simulation and interview. Examiners' consents to digitally record their comments were obtained verbally before the examination simulation. Their identities were protected in this study by using code names as shown in Table 3-3. The discussions relating to the examiners' comments would be anonymous. The code names were used only when necessary, such as to illustrate the between-examiner differences.

3.2.2 Student examination recordings and examination task

Interpreting examination recordings (English into Chinese) of five postgraduate students were selected for the main study. All five students were female. They were given code names and the aliases of Ally, Beth, Cherry, Daisy and Eileen as shown in Table 3-4. At the time of the recording, four students had received over a year's simultaneous interpreting training at postgraduate level, but Ally only had just over six months' training. Mandarin Chinese is A Language of all students except Cherry, whose A Language is English. To protect the identity of students, during the examination simulation the students were only referred to by code names, i.e. A, B, C, D, E, and aliases were used when discussing their performances in this thesis.

Table 3-4 Student background information for main study

Student / Code	Course exam mark	A Language	B Language
Ally / A	52	Chinese	English
Beth / B	66	Chinese	English
Cherry / C	71	English	Chinese
Daisy / D	55	Chinese	English
Eileen / E	58	Chinese	English

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 marked performers according to the marks given in one interpreting examination. Pollitt and Murray suggested that "a reasonable range of proficiency was deemed necessary to make sure a scale would emerge from the Thurstonian judgements, but it was predicted that comparisons at a similar ability level would elicit more insights into the less obvious aspects of rater behaviour" (1996: 79). This was also observed in the pilot study when examiners

considered more carefully in comparing the two close pairs of S2-S4 and S1-S5. In the main study, three students – Ally, Daisy and Eileen – had marks at the lower band of 50s and closer ability levels to each other (see Table 3-4). It was hoped that this would elicit more insights from the examiners when they compared the student performances.

The examination recordings for study were selected from a digital media archive of interpreting examinations made on a regular basis for teaching and research purposes, of which students' consent were obtained in class when attending the study course. All examination recordings were in digital *video* format to be played back on a laptop computer. The students' Chinese interpretations were 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 in the booth from the video recordings, and monitor both the target and source languages simultaneously, which is typical practice when assessing simultaneous interpreting.

The English source speech in the selected examination task was a keynote speech in a business conference. Based on the feedbacks of the pilot study, the part selected for the main study was the first three minutes of the speech[17], which was a general opening remark to introduce a company (see Appendix A).

3.2.3 Data recording and storage

Participant examiners' comments were recorded in digital audio files, stored on a hard drive, and transcribed into text documents for analysis. However, examiners' decisions of the paired comparisons and the results of the overall judgement ranking needed to be recorded on paper on-site by the researcher. Based on the experience in the

[17] The exam recordings are used as stimulus materials in this study. The study focus is on the participant examiners' judgements on the first 3 minutes of students' performances regardless if the students "get in the flow" or not.

pilot study, a worksheet and a simple marking table (see Appendix C) were designed to ensure fast and accurate data recordings without interrupting the examiners and delaying the examination simulation process.

The worksheet was used per examiner per sheet. It recorded the winners of the ten paired comparisons and the overall rankings, which was completed on-site by the researcher. The final overall marks given by the examiners at the end of each examination simulation were also recorded in the worksheet. In the main study, and the recording of the final overall marks from the marking tables onto the worksheet, and the ranking point conversion, were done after the examination simulation. So the examination process would not be interrupted (see 3.2.4).

3.2.4 Procedures of the main study

Basically the main study procedures were similar to those in the pilot study (3.1.1). The examination simulations and interviews were conducted with one examiner per session. So there were no between-examiner discussions; the data collection in this study was focused on individual examiners. This is because, as argued when setting the research aim and objectives in Chapter 1 (1.3), only when we know more about how individual examiners form their opinions before entering discussions with other examiners, can we better understand how they interact with each other in the interpreting examinations that are judged by a panel of examiners.

Through out the examination simulation and interview process, the researcher played a facilitator's role, administering the examination simulation, prompting the examiners to make paired comparisons (including asking simple questions to clarify vague comments, if any), recording the judgement results and the verbal comments of the examiners. To ensure that the same procedures were followed with every participant

examiner, procedural guidelines were written (see below) based on the experiences in the pilot study. This section explains the revised study procedures.

3.2.4.a Getting started – brief the examiners

The examination simulation started with a briefing on the purpose and general procedure of the study. To ensure standardised briefing of the examiners, I followed a written guideline as bellow to give the briefing session:

1. **Explain to the subject examiner the overall procedure of this experiment / interview.**
2. **Give the speech script to the examiner and play the source speech video clip. Let the examiner listen to the speech once when reading the script. If the examiner prefers to listen only and take notes, make a note of this preference at the back of the worksheet.**
3. **Ask if the subject examiner has any questions about the speech content.**

The examiners were advised that their verbal comment would be digitally recorded with their consents. They were informed that the examination was of postgraduate students who were in their second year of study on simultaneous interpreting in a two-year MA programme, and that the examination was a mid-term examination rather than a professional interpreting examination. They were told the language direction of the interpreting task, and no further detail of the students' backgrounds was revealed.

The examiners were asked *not* to give marks when making comparisons of the students' interpreting performances. This was to ensure the validity and effectiveness of the data collection. If they gave a mark to each student, it would be more likely that they simply remembered each student's mark in the later comparisons, paying less attention to the details of the differences in the students' performances for commenting.

Before listening to the students' performances, the examiners could read the speech script and listen to the speech from an audio recording, to familiarise themselves with the speech's content. Some examiners preferred just listening and taking notes of the

speech's contents without first reading the script, in which case the researcher would make a note of this preference at the back of the worksheet.

Since the purpose of the study is to investigate individual examiner's normal judgement process in interpreting examinations, their preferences were respected and they were not asked to change their normal assessment practice, nor were they given any assessment criteria or rating scale for the simulated examination in this study.

3.2.4.b Paired comparisons and commentary recording

After the briefing, the examination simulation began. This part broadly followed the procedure reported in the pilot study (3.1.1). The main difference was that worksheets and procedural guidelines were used in the main study so the data were recorded in a more systematic manner. The procedure followed the guidelines below:

1. **Give five copies of speech script to the subject examiner – one copy per student interpreter for examiner note-taking. Examiners could decide to use or not to use the script for note-taking when listening to the students' performances. Make a note of the examiner's script preference at the back of the worksheet.**
2. **Play student recordings A and B. Then stop to ask the subject examiner which one is better. Ask the examiner if reviewing the recording is needed. If yes, play the requested recording again. Then stop to ask which one is better again.**
3. **Start audio recording: ask the subject examiner to comment on the pair's performances. Questions to ask:**
    i. What are the pair's main differences?
    ii. Can you give the main reason(s) why this student is better, or worse?
4. **Stop audio recording when the subject examiner finishes commenting. The researcher writes down the code of the "winner" in the worksheet. Then play the next clip.**
5. **Repeat Steps 2 to 4. Play the remaining clips C, D and E, etc. until completing the comparisons of the ten pairs.**
6. **After comparing the ten pairs, ask the subject examiner to give their overall ranking of the five student interpreters' performance (start audio recording if more comments are given). Researcher writes down the rankings in worksheet.**

During the paired comparisons, the examiners were asked to focus on the pair of performances being compared and disregard the other performances. Whenever

examiners referred to a performance that was not currently being compared, I would stop the comment and remind the examiner to focus on the pair under comparison.

In keeping as much as possible to the normal practice of interpreting examination, i.e. student interpreters' performances are only watched or heard once in the examination process, each student's recording was normally played only once here in the study process; this could also save time in the simulation examination. If the participant examiner felt the need to review a previous recording so that they could compare with a different student's performance, the request was allowed, which is also what might happen in the real world when examination recordings are available to examiners for review purposes. In cases where such request was made, I would then make a note of the examiners' review situations on the worksheet for reference in the data analysis. Sixteen examiners felt no need to review and just carried on, while fourteen examiners made the request themselves to review, but with various frequencies and student performances.

At the end of the examination simulation, while the examiner had a fresh memory of the students' performances in mind, each examiner was asked to give an overall mark to each student's performance on a marking table; one student per marking table. The examination simulation ended when the marks were given to all five students.

3.2.5 Data analysis – a multi-strategy approach

The results of the collected data can be divided into two categories: (1) the quantitative data of the ranking points from the paired comparisons, overall judgements and marks, and (2) the qualitative data of the examiners' comments on the student interpreters' performances and why the winners were chosen. As seen in the research model (Figure 3-1), these two categories of data are to be cross-examined to investigate

how the examiners assess students' simultaneous interpreting performances. The strategy to report the study data is to firstly present and analyse the quantitative data in Chapter 4, i.e. the study on the examiners' reliability, followed by Chapters 5 and 6 that present and analyse the qualitative data, including the coding process; these two chapters aim to explore in detail the examiners' use of assessment criteria and their assessment behaviours.

The main study followed the same approach as in the pilot study to convert the paired comparison results into ranking points (3.1.2.a). The converted ranking points were then subjected to various statistical treatments, such as Thurstone's scaling technique and cluster analysis, to explore the examiners' judgement patterns. These statistical methods will be presented in detail in Chapter 4 as we explore the quantitative data to analyse the examiners' judgement consistency level, and to identify an analytical framework to cross-check their judgement patterns with the qualitative data, i.e. the examiners' paired-comparison comments.

Like in the pilot study, the examiners' comments were transcribed into text. The main study followed the coding practices of the Grounded Theory (see 2.5.4.c) to analyse this data qualitatively, examining the transcription of examiners' comments line-by-line, breaking them down and comparing them, and from these coding processes emerge concepts; similar concepts then can be further grouped into categories. Detailed illustrations of this qualitative data analysis approach will be presented in Chapter 5 after the quantitative data analysis.

In short, this research study combines the quantitative and qualitative methods in the hope of obtaining a better picture of the judgement process of the examiners in the interpreting examinations.

CHAPTER 4

Examiner Reliability Study

– a framework for assessment criteria analysis

4.0 Introduction

This chapter presents and analyses the quantitative datasets of the main study, comprising the results of the paired comparisons, overall judgement rankings, and overall marks. Coupled with the aim and objectives of this research study, this chapter intends to (1) identify and determine if the examiners have different judgement patterns, (2) to discover whether interpreter examiners and non-interpreter examiners can achieve consistent judgements, and then (3) to build an analytical framework for the qualitative data analysis of the examiners' use of assessment criteria.

In analysing the quantitative datasets, the first action was to investigate the overall consistency levels of the thirty examiners as a group. A one-way analysis of variance (ANOVA) was performed to establish whether the examiners separated the five students according to their interpreting performances (4.1). The datasets were then explored and analysed by using other statistical methods, including Cronbach's alpha for reliability and consistency estimations (4.2), and cluster analysis for grouping the subject examiners according to the consistency of their rankings of the student interpreters (4.3). Other explorations of the quantitative data were also carried out. The backgrounds of the examiners and their rankings of students' performances were examined on the basis of the results of the cluster analysis (4.4); Thurstone scales were also used to help the analysis during the data exploration process. Finally, the analytical framework based on the findings of this chapter is explained (4.5).

4.1 Results of the examiners' judgements as a group

This section presents the results of the paired comparisons, overall judgement rankings and overall marks. Using line graphs of the datasets, a cursory analysis of the results was carried out to identify initial patterns or any noticeable differences or similarities among the subject examiners' judgements (4.1.1). Then, some statistical analysis, i.e. ANOVA and the Thurstone scales were carried out with two objectives. The first objective is to identify whether the thirty examiners had significantly separated the five students' interpreting performances (4.1.2), which is to establish the usefulness of the stimulus materials and the examiners' judgement results. If the examiners as a group were unable to distinguish the five interpreting performances, it implies either that the five students performed at an equal level, or that the reliability level of examiners' judgement cannot be established. The second objective, then, is to find out the examiners' reliability level as a group when assessing the five students.

4.1.1 Results of paired comparisons, overall judgements, and overall marks

Each examiner's decisions on the students' paired comparisons were recorded in a ranking order worksheet (Table 3-5.a). The results were then converted into ranking points by using the method explained in 3.1.2.a. Table 4-1 presents the ranking points of the three assessment methods: paired comparison (hereafter PC), overall judgement (hereafter OJ), and overall marks (hereafter OM).

From examining the columns in Table 4-1, it appears that many examiners, if not most, were in agreement that Beth and Cherry, who received more high ranking points, were better than the other three, and that Ally, who received more low ranking points,

Table 4-1 Ranking points and overall marks

A: Ally **B:** Beth **C:** Cherry **D:** Daisy **E:** Eileen

Examiners	Ranking points & Overall marks														
	Paired comparison					Overall judgment					Overall mark				
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
R1	1	3	5	2	4	1	4	5	2	3	58	62	86	70	72
R2	2	4	3	1	5	2	3	4	1	5	65	75	80	40	90
R3	1	4	5	2	3	1	4	5	2	3	60	78	80	59	65
R4	3	4	5	1	2	3	4	5	1	2	77	84	88	69	80
R5	2	5	4	1	3	2	5	4	1	3	65	76	79	57	55
R6	1	5	4	2	3	1	4	5	2	3	59	81	80	65	59
R7	1	4	5	2	3	1	4	5	2	3	69	73	78	73	75
R8	2	4	5	1	3	3	4	5	1	2	75	85	90	60	60
R9	5	5	3	3	1	5	4	2	3	1	72	75	68	65	58
R10	4	5	3	2	1	4	5	3	2	1	75	78	65	65	60
R11	1	4	5	3	2	2	4	5	3	1	70	79	85	75	73
R12	3	4	5	2	1	3	5	4	2	1	55	73	68	72	70
R13	2	4	5	1	3	2	4	5	1	3	65	85	85	60	70
R14	1	5	2	4	3	1	5	3	4	2	55	80	65	68	69
R15	1	4	5	2	3	1	4	5	3	2	68	78	83	75	75
R16	1	2	5	3	4	1	2	5	3	4	20	65	70	60	67
R17	4	3	5	2	1	3	4	5	2	1	75	80	83	65	65
R18	1	3	4	5	2	1	3	5	4	2	55	60	68	67	58
R19	2	5	4	3	1	2	5	4	3	1	70	75	73	72	68
R20	1	5	3	1	3	1	5	3	2	4	55	65	58	56	60
R21	3	4	5	1	2	3	4	5	2	1	70	85	85	65	50
R22	1	5	4	3	2	1	5	4	3	2	70	88	85	80	78
R23	2	5	4	1	3	2	4	5	1	3	65	75	79	70	55
R24	1	5	3	4	2	1	5	4	3	2	50	90	85	85	80
R25	1	5	4	2	3	1	5	4	2	3	58	80	72	68	60
R26	2	3	5	4	1	2	3	5	4	1	75	80	85	70	67
R27	1	4	5	2	3	1	4	5	3	2	58	79	82	62	65
R28	2	3	5	1	4	2	3	5	1	4	75	75	85	60	78
R29	1	3	5	3	3	1	3	5	3	3	54	75	88	75	75
R30	1	3	5	2	4	1	2	5	3	4	55	68	75	65	70

Ranking Points: 5 – 1st place, **4** – 2nd place, **3** – 3rd place, **2** – 4th place and **1** – 5th place

was the worst interpreter. However, it is also obvious that many examiners' judgements varied from one another. Although some random transposition of ranks is inevitable when it comes to subjective judgement like those in this study, there appears to be bigger variations that raise concerns. For example, examiners' judgements on Ally, Daisy and Eileen ranged from the worst to the best, while Beth and Cherry ranged from the best to the 4th. Browsing across the rows on Table 4-1, it is also noticeable that the individual examiners' judgements were not always consistent as individual examiner's rankings changed between the PC and the OJ assessment methods.

Figures 4-1.a and 4-1.b are the line graph presentations of PC and OJ rankings, which make evident the sometimes wide variations between the examiners' judgements. The lines should have appeared roughly level with minor variations only if the between-examiner judgements were similar. Nevertheless, the line graphs also show similar fluctuating patterns of the examiners' judgement as a group between the PC and OJ assessment methods. Beth and Cherry are mostly in the upper part of the graphs with some dips, while Ally is often at the bottom with some upward spikes. Daisy and Eileen are both mostly in the lower half of the graphs, touching the bottom more often than reaching the higher end of the ranking points.

Figure 4-1.c is the line graph presentation of the overall marks, which does not at first sight look similar to the other two line graph patterns. However, when the overall marks are converted into ranking points, i.e. disregarding the actual distance between the marks, and then plotted into the line graph shown in Figure 4-1.d, the OM ranking point line graph resembles to the PC and OJ line graphs.

A cursory analysis of the patterns of these line graphs (Figures 4-1.a, 4-1.b and 4-1.d) shows that the examiners as a group, despite the fluctuations between them individually, seem to be judging the five students similarly in terms of the ranking patterns among the three assessment methods.

Figure 4-1.a Line graph – paired comparison (PC) rankings

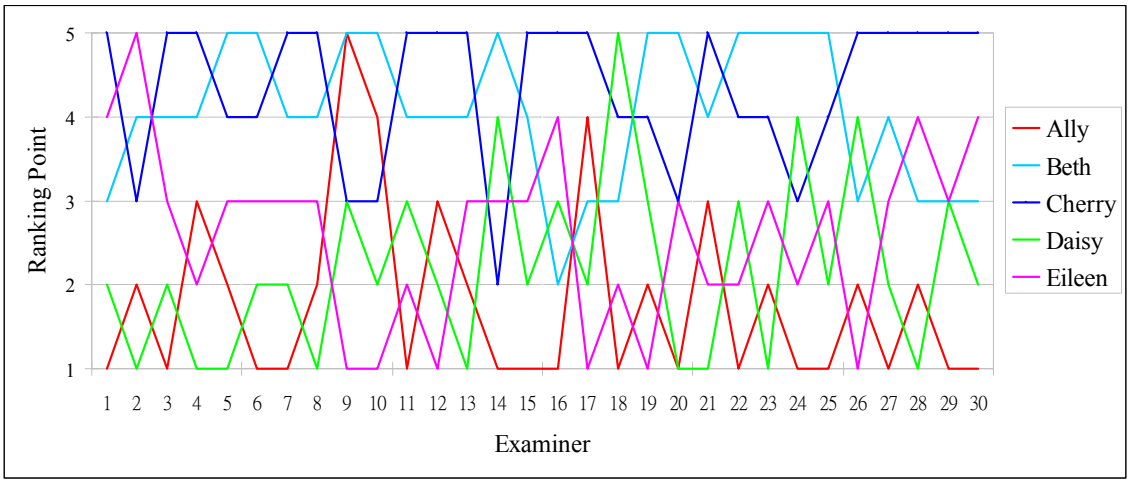


Figure 4-1.b Line graph – overall judgment (OJ) rankings

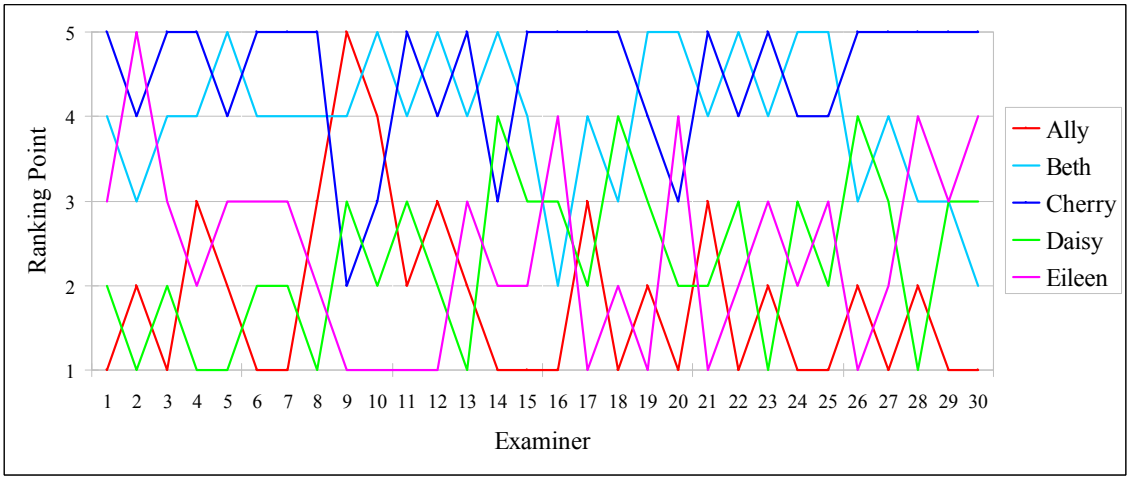


Figure 4-1.c Line graph – overall marks (OM)

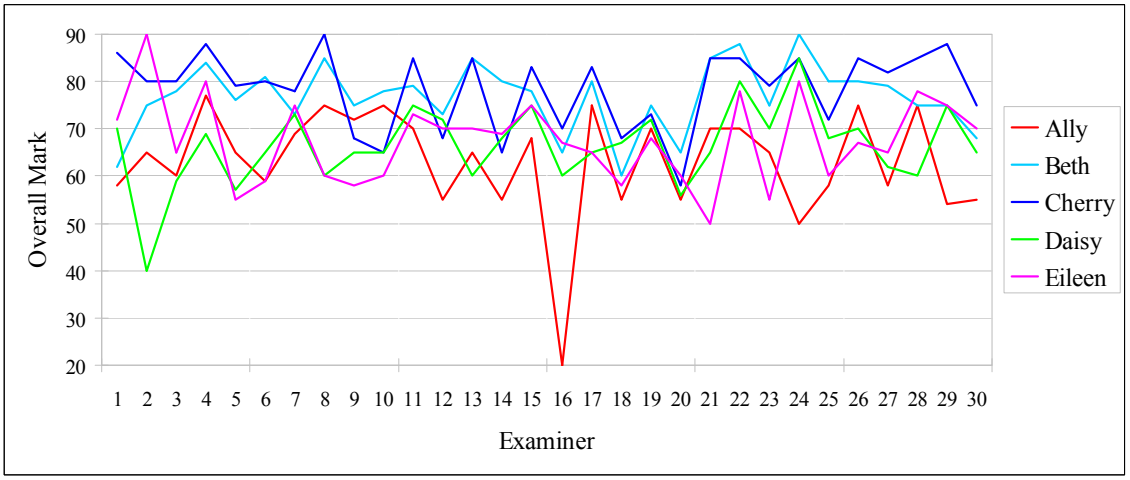
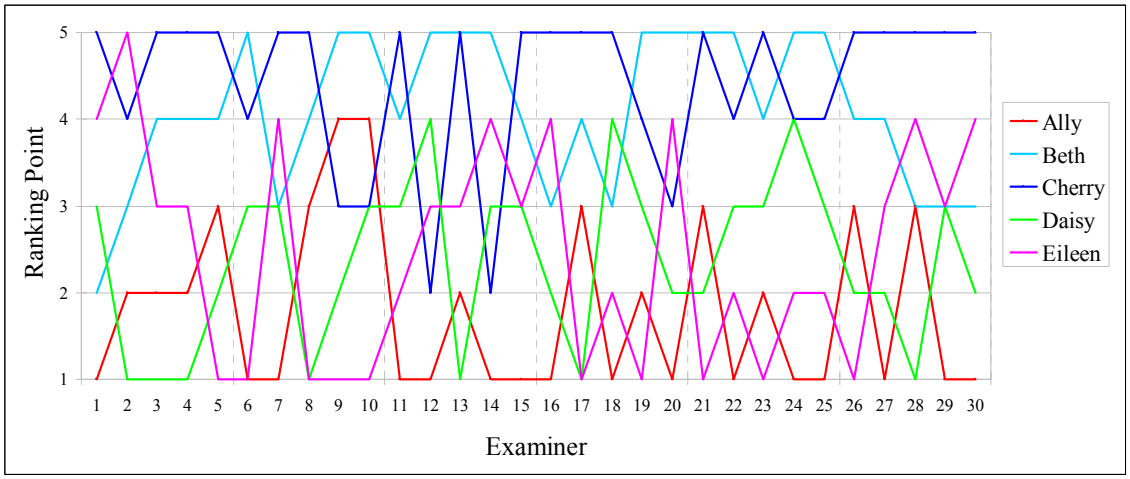


Figure 4-1.d Line graph – overall marks (OM) converted into ranking points



4.1.2 Examiners' reliability as a group

In this section, the statistical analyses will be carried out to see if the separation of the five students is significant in terms of their ranking points, and to see from the statistical point of view if the thirty examiners as a group assess the students consistently when using the three methods, i.e. PC, OJ and OM.

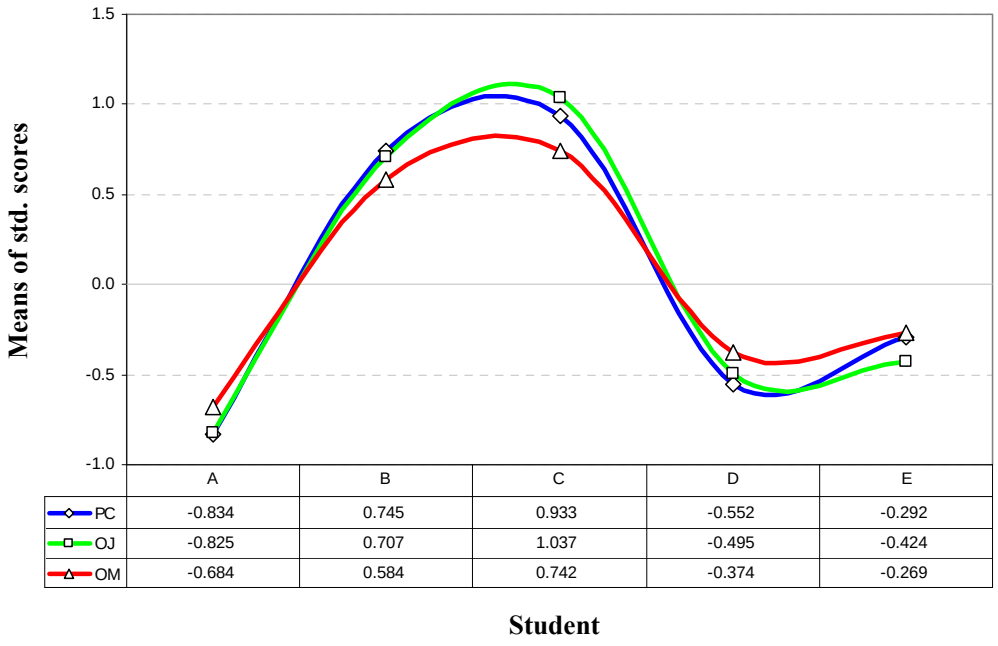
4.1.2.a Results of one-way ANOVA

Using the datasets in Table 4-1, a one-way ANOVA analysis is performed. The ranking points and overall marks are transformed into standardised scores[18] for analysis. ANVOA is used to test the null hypothesis that several means are equal, or not significantly different. If the one-way ANOVA shows that the student datasets are significantly different, it shows that the examiners as a group have managed to separate the five students according to their interpreting performances.

[18] Standardised scores are also called *z*-scores. In statistics, a standard score indicates how many standard deviations an observation or datum is above or below the mean. It is often used to statically "standardise" different set of scores, in this case, the ranking points and overall marks, for comparisons on an equal footing in statistical terms (StatSoft., 2006). See Appendix E.

The ANOVA results of the three assessment methods show that all three *p* values are less than 0.001 (PC: $F(4,154) = 37.097$, $p < 0.001$; OJ: $F(4,154) = 42.524$, $p < 0.001$; OM: $F(4,154) = 16.792$, $p < 0.001$), indicating that the five students are highly significantly different in terms of their rankings and marks of the three assessment methods (see Appendix F). Therefore, we can very confidently say that in spite of the judgement variations observed in the line graphs, the subject examiners as a group are successful in separating the five students' interpreting performances when using the three assessment methods.

Figure 4-2 Line graph of means of standardised scores – PC, OJ and OM



	A	B	C	D	E
PC	-0.834	0.745	0.933	-0.552	-0.292
OJ	-0.825	0.707	1.037	-0.495	-0.424
OM	-0.684	0.584	0.742	-0.374	-0.269

The means of the standardised scores are also plotted as shown in Figure 4-2. The three line patterns that represent the three assessment methods are identical. This echoes the earlier observation from Figures 4-1 that the subject examiners as a group assessed the students in a similar way in terms of ranking results when using the three assessment methods. In other words, each of the three ranking or assessment methods are equally effective, though the fact that they were used together in this study means the certain memory effects cannot be ruled out.

4.1.2.b Three Thurstone scales

An alternative way of looking at these data is given by Thurston scales. Three Thurstone scales (hereafter T scales) were produced based on the datasets from the three assessment methods shown in Table 4-1. The T scales can be regarded as interpreting proficiency scales to show the five students' relative positions according the perceptions of thirty examiners of how well they performed. The left of the scales indicates better performance and the right part of the scale indicates worse performance.

Figure 4-3 Thurstone scales of interpreting proficiency

a. Paired comparison

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

b. Overall judgement

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

c. Overall mark

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)*

On each T scale, the order of the five students is the same with Cherry and Beth ahead, followed by Eileen, Daisy, and Ally. This is in line with the previous observations (Figures 4-1 and Figure 4-2) where the thirty examiners as a group judged the five students consistently on 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 (Table 3-4).

The only noticeable difference among the three scales is the gap between Eileen and Daisy. The gap between the two students' positions appears wider on the PC T scale (Figure 4-3 a) than on the other two scales (Figures 4-3 b and 4-3 c). 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 in the paired comparison, the larger gap in 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 terms of 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.

4.2 Using Cronbach's alpha to examine consistency levels

In the pilot study we have reported using Cronbach's alpha to estimate the consistency level of two Thurstone scales (3.1.2). In this section, Cronbach's alpha (a statistical correlation coefficient) is calculated again to analyse the consistency levels of the examiners' judgements, i.e. the ranking points, in the main study. This time, the alpha's intra-class correlation coefficients (ICC) are used to cross-examine and analyse examiners' consistency levels (1) between and within the three assessment methods (4.2.1), and (2) according to the examiners' different characteristics and backgrounds within each of the three assessment methods (4.2.2 and 4.2.3).

ICC is an ANOVA-type statistical method in which the examiners' rankings, or marks, are responses. There are two sources of variation in the current study's examination simulation. One source is the student interpreters' performances, which are a random sample from a large pool of performances; the other source is the examiners, who are a random sample from a large pool of examiners. Therefore, a two-way random effects model is selected for the ICC computation.

ICC provides two types of measures: average measures and single measures. The value of the average measures is the averaged ICC scores of all the subject examiners, i.e. the consistency of a group. The value of the single measures indicates the reliability level if only one subject examiner is used, i.e. the consistency of an individual relative to the group. The ICC computation is set to check the consistency level of the scales (i.e. ranking points) with a 95% confidence level.

4.2.1 Cronbach's alpha (ICC) – all examiners

4.2.1.a Consistency between assessment methods

Table 4-2 Cronbach's alpha for three T scales – intra-class correlation coefficient

All examiners' judgments		Intra-class correlation	95% confidence interval	
			Lower bound	Upper bound
Three T scales: PC, OJ, OM	Single measure	0.99	0.97	0.99
	Average measure	0.99	0.99	1

This section will first ascertain statistically if the examiners assess consistently between the three assessment methods. The T scale values are used for the alpha ICC calculation to estimate the reliability level of the three T scales. As shown in Table 4-2, the alpha, i.e. the *average measure*, is excellent at 0.99. This high alpha means that the three T scales are highly consistent with one another. The *single measure* ICC is also excellent at 0.99, which means that each T scale can be used individually and still achieves an excellent consistency level. The high scores across the lower bound and upper bound also indicate that the three scales are highly consistent with each other when measuring the five students' interpreting proficiency. The above result statistically confirms the previous observations in 4.1 that the thirty examiners as a group assessed the students consistently across the three assessment methods.

4.2.1.b Consistency between examiners

However, fluctuations in the line graphs (Figures 4.1) indicate consistency problems between the examiners. The alpha ICC can estimate how consistently and reliably the examiners exercise their judgement *within* each assessment method, i.e. between the examiners. This time, the ranking points and overall marks given by the examiners, i.e. their judgement, will be used for the alpha ICC calculation. The results are shown in Table 4-3.

Table 4-3 Cronbach's alpha for all examiners – intra-class correlation coefficient

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

Firstly, all three average-measures alphas are above 0.9 and the average-measures ICC scores between the lower bound and the upper bound are also excellent (mostly >0.9). These indicate that the examiners' group consistency levels can be considered as excellent, which is in line with the results in 4.2.1.a, though overall marking (0.95) appears to be slightly less consistent than the other two methods (0.97) (*cf.* 4.1.2.b).

However, the low values of the single-measures ICC scores (0.49, 0.52, 0.41) suggest poor and unacceptable consistency levels of *individual* examiners' judgements when assessing the students. Furthermore, the range in which the single-measures ICC scores fluctuate is wide between a low lower bound (<0.3) and a high upper bound (>0.8). These results suggest that if just one individual examiner is used in the interpreting examinations, the judgement results would not be consistent and reliable, which reflect the observed between-examiner fluctuations in Figures 4-1.

So far, it has been confirmed statistically that the thirty examiners as a group assessed the students with a good consistency level. However, it is impractical to use thirty examiners in an interpreting examination to achieve a consistent result. As the single-measures ICC shows that individual examiners are unlikely 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 judgement. For example, can a smaller group of examiners of the same background achieve a satisfactory reliability level? In the sections below, the examiners' backgrounds will be checked to answer this question.

4.2.2 Cronbach's alpha (ICC) – according to examiners' backgrounds

Table 4-4 Cronbach's alpha (ICC) – according to examiners' background

All examiners / Intra-class correlation coefficient	Paired comparison		Overall judgment		Overall mark		Number of Examiners
	Single measures	Average measures	Single measures	Average measures	Single measures	Average measures	
Interpreters	0.43	0.93	0.48	0.95	0.39	0.92	19
Translator	0.72	0.97	0.68	0.96	0.48	0.91	11
Teaching SI	0.46	0.92	0.55	0.94	0.43	0.91	13
Using script	0.58	0.97	0.63	0.97	0.46	0.95	21
No script	0.43	0.87	0.34	0.82	0.36	0.83	9

Table 4-4 presents the alpha ICC results of the three assessment methods according to the examiners' backgrounds. The average measures, i.e. the alphas for group consistency levels, for all of the three methods are above 0.9 for examiners of all backgrounds, except those who did not use scripts (0.83), which can still be considered as good. In the PC method, it is the translator examiners (0.97) and those who used scripts (0.97) that have the highest group consistency levels. In fact, the examiners who used scripts are consistently among the highest across all three assessment methods (0.97, 0.97, 0.95).

Another interesting finding is that when using PC and OJ methods, translator examiners are more consistent (0.97, 0.96) than the interpreter examiners (0.93, 0.95). This finding is largely in line with the results observed in the pilot study where non-interpreter examiners achieved a better consistency level than interpreter examiners (Wu, 2010). Here, only in giving overall marks did the interpreter examiners assess the students with a slightly higher consistency level (0.92) than the translator examiners (0.91). Nevertheless, these differences are too small to be statistically significant.

Looking at the single-measures ICC in Table 4-4, however, two thirds are below 0.5, and none are above 0.5 using the OM method. This suggests that regardless of their backgrounds the consistency level would be unacceptable if only one individual

examiner is used for overall-marking. The only assessment method where the individual examiners can be considered as judging at an acceptable consistency level is the PC method by translator examiners (0.72). However, they deteriorate to a questionable level for the OJ method (0.68), then to unacceptable for the OM method (0.48). Even so, individual translator examiners still assessed students more consistently across the three methods than the interpreter examiners, whose consistency levels were all unacceptable (0.43, 0.48, 0.39).

In summary, as a group translator examiners in general have a higher consistency level than the interpreter examiners; examiners who used the exam speech scripts appear to be more consistent than those who did not use the scripts. As individuals, only the translator examiners achieved an acceptable level of consistency, though only when using the PC assessment method.

4.2.3 Cronbach's alpha (ICC) – translator vs. interpreter examiners

Given the multiple characteristics of the examiners (see Table 3-3), this section further compares the translator and the interpreter examiners in sub-groups.

4.2.3.a Translator examiners

Table 4-5 Cronbach's alpha (ICC) – cross-examining translator examiners

Translator examiners / Intra-class correlation coefficient	Paired comparison		Overall judgment		Overall mark		Number of examiner
	Single measures	Average measures	Single measures	Average measures	Single measures	Average measures	
Using script	0.78	0.97	0.74	0.96	0.63	0.94	9
No script	0.50	0.67	0.60	0.75	0.46	0.63	2

Table 4-5 shows that those translator examiners who used scripts have higher scores in both single- and average-measures ICC. The low number of No-script

examiners is low here, and the statistic results may not be rigorous for generalisation. Nevertheless, the results may broadly demonstrate that those examiners who use scripts are more consistent than those who did not use scripts. Using the script *may* help the translator examiners to make judgements that are more reliable (*cf.* No SI teaching + No script in Table 4-6 below).

4.2.3.b Interpreter examiners

Table 4-6 Cronbach's alpha (ICC) – cross-examining interpreter examiners

Interpreter examiners / Intra-class correlation coefficient	Paired comparison		Overall judgment		Overall mark		Number of examiner
	Single measures	Average measures	Single measures	Average measures	Single measures	Average measures	
Teaching SI	0.46	0.92	0.55	0.94	0.43	0.91	13
No SI teaching	0.33	0.74	0.31	0.73	0.19	0.58	6
Using script	0.46	0.91	0.58	0.94	0.38	0.88	12
No script	0.45	0.85	0.34	0.79	0.45	0.85	7
Teaching SI + Using script	0.43	0.86	0.56	0.91	0.37	0.83	8
Teaching SI + No script	0.61	0.89	0.57	0.87	0.52	0.85	5
No SI teaching+ Using script	0.40	0.73	0.55	0.83	0.28	0.60	4
No SI teaching + No script	0.94	0.97	0.80	0.89	0.94	0.97	2

In Table 4-6 in respect to all three assessment methods, the examiners with SI teaching background as a group (i.e. the average measures) are more consistent (all >0.9) than those without SI-teaching background (all <0.75); in particular, interpreters without SI-teaching background as a group have a poor consistent level (0.58) when using the overall marking method. As individuals (i.e. the single measures), examiners' consistency levels in both groups are unacceptable (mostly <0.5), though those who teach SI (0.46, 0.55, 0.43) are still better than those who do not teach (0.33, 0.31, 0.19).

For script usage preference as a group (i.e. the average measures) across the three methods, it is also clear that interpreter examiners who use scripts (0.91, 0.94, 0.88) are more consistent than those who do not use scripts (0.85, 0.79, 0.85). However, the gap here is not as great as in the SI-teaching background, especially when using the OM

method (using-script 0.88, no-script 0.85), which is still reasonably good. As individuals (i.e. the single measures), the examiners' consistency levels are mostly unacceptable (<0.5), using or not using script. When giving overall marks, however, it is a surprise to see that individual interpreter examiners who *do not* use scripts are slightly more consistent (0.45) than those who use scripts (0.38), though both are still unacceptable in terms of consistency level. This is similar to the situation in the SI teaching vs. non-teaching comparison as observed above.

The ICC results in the second half of Table 4-6 show further analysis between the examiners with and without SI-teaching background in relation to their script-use preferences. Very interestingly, when the interpreter examiners are separated in such two groups according to the teaching backgrounds, those who *do not* use script appear to assess the students more consistently than those who use script, as a group *and* as individuals across the three methods, except the SI-teaching group in OJ method (using-script 0.91, no-script 0.87). In particular, the two interpreter examiners without SI-teaching have achieved very good and excellent consistency levels across the three assessment methods without using a script, both as a group (0.97, 0.89, 0.97) and as individuals (0.94, 0.80, 0.94), which is impressive considering the other single measures are all relatively low at around the 0.5 level.

Compared to the translator examiners, this indicates a unique characteristic of the way interpreter examiners assess student interpreters. Interpreter examiners may be more used to assess without using a script. Being an interpreter, they rely more on listening than on reading when receiving and processing messages. Nevertheless, as the number of subject examiners is small, these results may be within the range of random variations, especially for the No-SI-teaching and No-script-use comparison. The lower and smaller gaps between the ICC scores here, however, may imply that the examiners' backgrounds play a less important role in their judgement patterns.

4.2.4 Summary discussion

From the above findings, it appears that overall marking is less consistent among the three methods when assessing the student interpreters (4.1.2.b, 4.2.1.b). We also found that using scripts may help translator examiners more in improving consistency levels than it helps the interpreter examiners, and experience in teaching SI seems to benefit interpreter examiners more in achieving a higher consistency than does using speech scripts (4.2.2, 4.2.3.b). One possible intuitive explanation is that using a script may help translators with no SI experience more, especially because they can then fall back on their experience of judging from written text. Whereas, for those with simultaneous interpreting experience, familiarity with simultaneous interpreting and their experience of its assessment experience from teaching overshadow the script factor as interpreters are more used to receiving messages by listening to them.

The consistently low single measures ICC, however, suggest that there may be more attributes involved in judgement consistency than the pre-defined predictors. For example, the two interpreter examiners who do not teach SI and do not use script still achieve highly reliable ICC scores of both single and average measures. This implies that there may be factors other than the teaching background and the use of script operating in relation to the reliability level of assessing simultaneous interpreting.

Therefore, more analysis of the examiners' judgements will be carried out to explore other possible groupings of the examiners, and in particular to attempt to identify more and less consistent types, or groups, of examiners.

4.3 Using cluster analysis to explore types of examiners

In 4.2.1, it was argued that it is impractical to use thirty examiners in an interpreting examination panel so it would be useful to know the minimum, or optimum number of examiners required to achieve a good consistent judgement result. Since no answer to this question could be drawn from the analysis of the examiners' background, in this section I would attempt to allocate the examiners to groups based ranking points alone using the cluster analysis method.

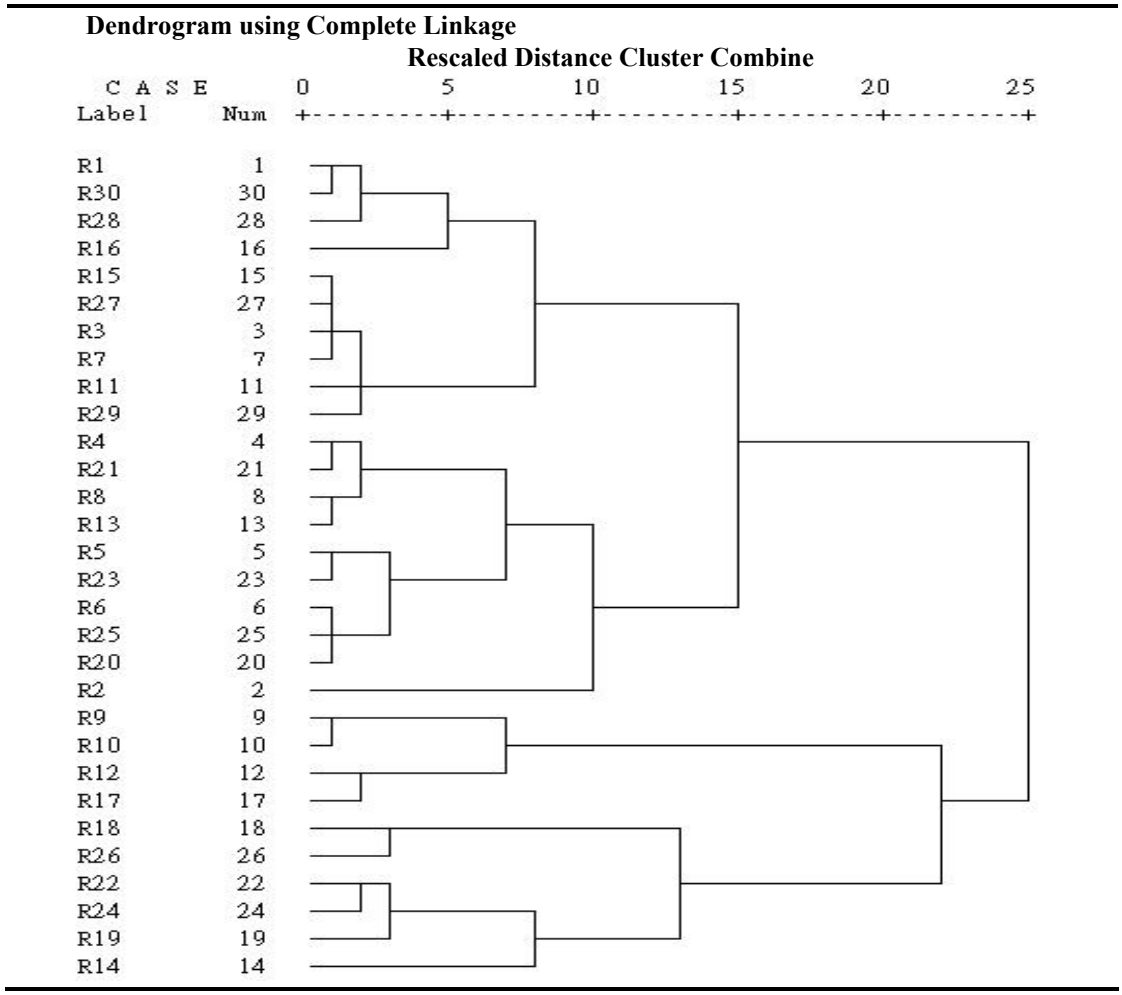
Cluster analysis is an exploratory tool for quantitative data analysis (StatSoft, 2006). It uses statistical software, and is basically a collection of algorithms that put items (in this case, the examiners) into clusters according to pre-defined similarity rules: the similarity association is maximal (in this case, the ranking point patterns) if the items belong to the same group and minimal otherwise. Thus, cluster analysis can identify structures in data items without providing an explanation, and is mostly used when the research is still in the exploratory phase (ibid). The results of the cluster analysis then could be used as a framework for further qualitative analysis (in the present study, examiners' comments) to find the reasons why some examiners are in the same group and others not. By doing so, it is hoped to answer this study's research questions.

The three ranking point datasets – PC, OJ, and OM – were analysed using SPSS software to perform cluster analysis. A typical cluster analysis output is a dendrogram[19], or a hierarchical tree plot, as seen in Figures 4-4.a, 4-4.b and 4-4.c.

[19] The cluster analysis dendrograms are generated by selecting the complete linkage method with squared Euclidean distance in the SPSS software.

4.3.1 Cluster analysis of all examiners – paired comparison

Figure 4-4.a Cluster analysis of all examiners – paired comparison



In this section, the dendrogram representing PC assessment method (Fig. 4-4.a) is used to illustrate how a suitable number of clusters may be extracted from the thirty examiners. On the left side of the figure, the case labels represent the thirty examiners; and on the top of the tree plot, there is a scale from 0 to 25, which can be used to identity the distances between clusters. At 0-distance, for example, all thirty examiners are separated, i.e. thirty items, but on moving toward the right, the examiners start to join into clusters. Take R1, R30, R28 and R16 for example, R1 and R30 first join together, then R28 joins this cluster at about distance 2; then R16 joins the R1-R30-R28

cluster at distance 5 to form a 4-member cluster, which combines with a 6-member cluster at about distance 8 to form a 10-member cluster. At distance 15, this 10-member cluster joins with another 10-member cluster, forming a 20-member cluster. Finally, to the far right at distance 25, all 30 examiners join together to complete the dendrogram.

To extract clusters of suitable size for analysis, a cut off point can be selected on the distance scale. For example, if the cut off point were selected at distance 20, three clusters would be extracted: a 20-member cluster, a 4-member cluster, and a 6-member cluster; whereas, with the cut off point at distance 14, the 20-member cluster separates into two 10-member clusters and four clusters will be extracted.

As the cluster analysis puts similar items together first, the farther to the right on the distance scale, the items (i.e. the examiners) become more disassociated, hence less consistent. Our aim is to identify groups of examiners that achieve higher levels of consistency. A suitable cut off point on the distance scale needs to be decided to extract the clusters to achieve this aim. So, the results from Cronbach's alpha ICC were used to help to decide the cut off point. A good cut off point, therefore, would be where the clusters first achieve a consistency level of more than 0.7 in the single measures ICC, which indicates that individual examiners in the same cluster can judge consistently at an acceptable level.

It was found that a cut off point at about distance 12 first achieved the aim of improving all single measures ICC to above the 0.7 level in the extracted clusters. Table 4-7.a presents the cluster membership of the examiners and their ICC scores. The average measures ICC are all excellent with only Cluster 4 slightly below the 0.9 mark, and the single measures ICC of the five clusters are all above 0.7 with Cluster 1 and 4 at 0.8 and above. These are much improved results if compared with those in Table 4-4 when the examiners are grouped according to their background. So, the five groups of examiners identified are internally consistent within each cluster.

Table 4-7.a PC cluster membership and ICC

Clusters	Examiners	Number of Examiners	Intra-class Correlation Coefficient	
			Single measures	Average measures
1	R1, R3, R7, R11, R15, R16, R27, R28,R29, R30	10	0.87	0.98
2	R2, R4, R5, R6, R8, R13, R20, R21, R23, R25	10	0.76	0.97
3	R9, R10, R12, R17	4	0.71	0.91
4	R18, R26	2	0.80	0.89
5	R14, R19, R22, R24	4	0.78	0.94

Following the method described above, results of the cluster analysis and the cluster memberships of the examiners for the other two assessment methods – OJ and OM – are presented in 4.3.2 and 4.3.3 below.

4.3.2 Cluster analysis of all examiners – overall judgement

Figure 4-4.b is the OJ dendrogram. As shown in the figure, the clusters differ from those in Figure 4-4.a. However, the cut off point to improve the single measures ICC above the 0.7 threshold is similar at just short of 12 on the distance scale, but here just four clusters are extracted. The results of the cluster membership and their respective ICC scores are presented in Table 4-7.b1.

The average measures are excellent at the 0.97 level for Clusters 1, 2 and 3, with Cluster 4 slightly below 0.9. The single measures ICC are also all much improved with Cluster 1 at 0.731 and three clusters above 0.8. Cluster 1 is a 14-member cluster but has the lowest single measures ICC score. If the cut off point is reduced to distance 10, for example, Cluster 1 can be separated into two clusters: one has 3 members (R16, R29, R30) and the other has 11 members, whose single measures ICC are 0.93 and 0.80 respectively. This will increase the total number of clusters to five all having single measures ICC at or above the 0.8 level as shown in Table 4-9.b2.

Figure 4-4.b Cluster analysis of all examiners – overall judgment

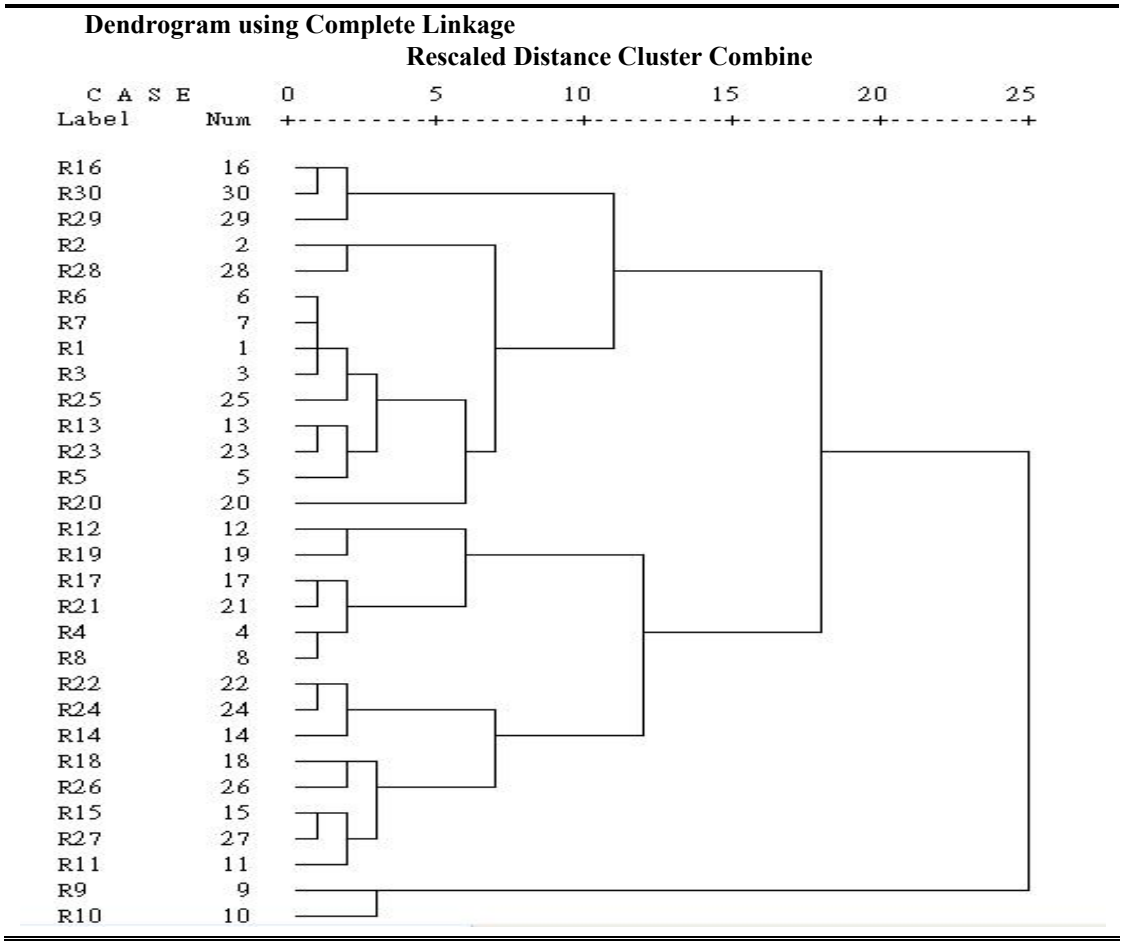


Table 4-7.b1 OJ cluster membership and ICC – 4 clusters

Clusters	Examiners	Number of examiners	Intra-class correlation coefficient	
			Single measures	Average measures
1	R1, R2, R3, R5, R6, R7, R13, **R16**,R20, R23, R25,R28, **R29**, **R30**	14	0.73	0.97
2	R4, R8, R12, R17, R19, R21	6	0.85	0.97
3	R11, R14, R15, R18, R22, R24, R26, R27	8	0.80	0.97
4	R9, R10	2	0.80	0.89

Table 4-7.b2 OJ cluster membership and ICC – 5 clusters

Clusters	Examiners	Number of examiners	Intra-class correlation coefficient	
			Single measures	Average measures
1	R16, R29, R30	3	0.93	0.98
2	R1, R2, R3, R5, R6, R7, R13, R20, R23, R25, R28	11	0.80	0.98
3	R4, R8, R12, R17, R19, R21	6	0.85	0.97
4	R11, R14, R15, R18, R22, R24, R26, R27	8	0.80	0.97
5	R9, R10	2	0.80	0.89

4.3.3 Cluster analysis of all examiners – overall mark

Figure 4-4.c presents the OM dendrogram, which is very different from the other two dendrograms (Figures 4-4.a and 4-4.b). The cut off point to improve the single measures ICC above the 0.7 threshold is at about 3 or 4 on the distance scale, which means that there can be as many as 12 clusters. This is not ideal for analysis because of the large number of clusters.

Therefore, it was decided to set the cut off point at distance 10 where five clusters are extracted just like the other two dendrograms, and the results are presented in Table 4-7.c. Items at distance 10 are less associated with each other than those at distance 3 or 4, which would make the alpha ICC scores lower. As shown in Table 4-7.c, Clusters 3 and 5 are less consistent than the other clusters in terms of their average and single measures ICC scores; for the single measures both clusters only achieve a 0.6 level.

For the OM assessment method, therefore, it is more difficult to extract the same number of clusters with a single measure ICC threshold at the 0.7 level, the level set for the two other assessment methods. This indicates that the overall marking method is potentially the least consistent of the three assessment methods for assessing student interpreters. Compared with the methods which rank students, it is more difficult to measure consistently how much better or worse than one another members of a group of students are by assigning a numeric mark to each of them, especially when the students' ability levels are similar.

The finding here is in line with the observations in the comparison of the three Thurstone scales (4.1.2.b) and the Cronbach's alphas analysis (4.2), in which the OM method appears to be less consistent than the PC and OM methods.

Figure 4-4.c Cluster analysis of all examiners – overall mark

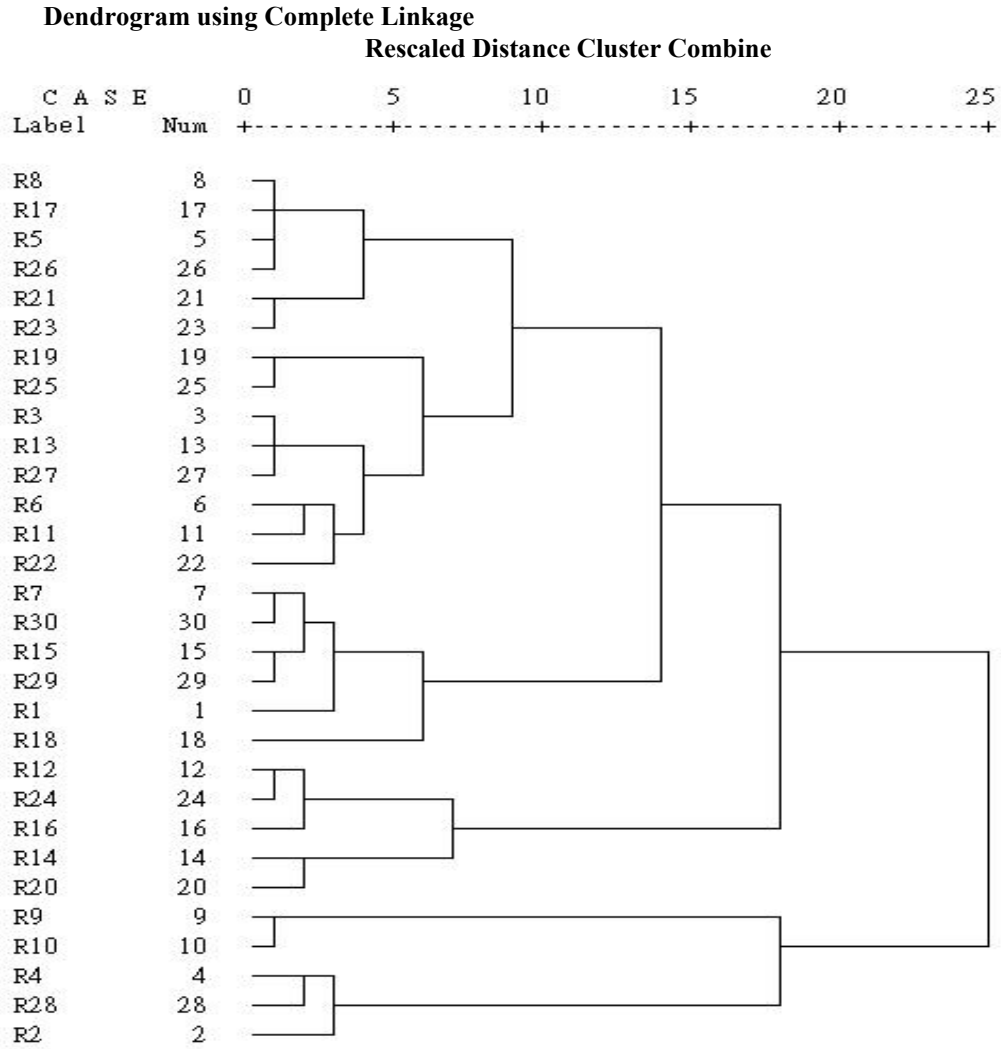


Table 4-7.c OM cluster membership and ICC

Clusters	Examiners	Number of examiners	Intra-class Correlation Coefficient	
			Single measures	Average measures
1	R3, R5, R6, R8, R11, R13, R17, R19, R21, R22, R23, R25, R26, R27	14	0.73	0.97
2	R1, R7, R15, R18, R29, R30	6	0.71	0.94
3	R12, R14, R16, R20, R24	5	0.61	0.89
4	R9, R10	2	0.94	0.97
5	R2, R4, R28	3	0.63	0.85

4.4 Pattern of examiners by cluster membership

In this section, the cluster membership extracted in the previous section is cross-examined with examiner background. The objective is to see if any new patterns emerge under the condition that the examiners assess the five student interpreters more consistently, that is, internally within each cluster.

Table 4-8 Cluster membership of examiners in the three assessment methods

✍: translator 🎧: interpreter 🗣: SI teaching 📄: use script

	Examiner Background			Cluster membership		
				Paired comparison	Overall judgement	Overall mark
R1	✍		📄	1	2	2
R2	✍			2	2	5
R3	✍			1	2	1
R4	🎧	🗣		2	3	5
R5	🎧	🗣		2	2	1
R6	🎧	🗣		2	2	1
R7	✍		📄	1	2	2
R8	🎧	🗣	📄	2	3	1
R9	🎧			3	5	4
R10	🎧			3	5	4
R11	✍		📄	1	4	1
R12	🎧	🗣	📄	3	3	3
R13	🎧		📄	2	2	1
R14	🎧	🗣		5	4	3
R15	✍		📄	1	4	2
R16	🎧	🗣	📄	1	1	3
R17	🎧	🗣	📄	3	3	1
R18	🎧	🗣	📄	4	4	2
R19	🎧		📄	5	3	1
R20	🎧	🗣		2	2	3
R21	🎧	🗣	📄	2	3	1
R22	🎧	🗣	📄	5	4	1
R23	🎧	🗣	📄	2	2	1
R24	🎧		📄	5	4	3
R25	✍		📄	2	2	1
R26	✍		📄	4	4	1
R27	✍		📄	1	4	1
R28	🎧		📄	1	2	5
R29	✍		📄	1	1	2
R30	✍		📄	1	1	2

Table 4-8 shows the examiners' cluster membership in the three assessment methods and the corresponding background of each examiner. Looking across the columns, it appears that examiners' cluster membership may change from one assessment method to another, showing inconsistent cross-method groupings. When cross-examining the examiners' background in Table 4-8, it is also difficult to see a clear relation between the backgrounds and the cluster memberships.

Nevertheless, interesting patterns emerge in terms of examiner backgrounds when the order of the examiners in Table 4-8 is re-arranged according to the cluster dendrograms as shown in Tables 4-9.a, 4-9.b and 4-9.c below.

4.4.1 PC clusters and the examiners' backgrounds

Table 4-9.a shows the examiners' cluster membership and their background in the order from the PC dendrogram (Figure 4-4.a). In general, within the individual clusters membership seems homogenous. PC Cluster 1 (hereafter PCC1) consists mostly of translators; PC Cluster 2 (hereafter PCC2) is mostly interpreters; PC Clusters 3 and 5 (hereafter PCC3, PCC5) are all interpreter examiners, while PC Cluster 4 (hereafter PCC4) consists of one interpreter and one translator.

As the clusters were extracted with the condition that examiners' single measures ICC needed to be above 0.7, the consistency level of individual examiners is considered acceptable within each cluster regardless of differences in the examiners' backgrounds. That is to say, the examiners' divergent professional backgrounds and use of script do not necessarily determine the consistency level of their judgements within a cluster. For example, although PCC1 contains mostly translators, the two interpreters in the same cluster share similar judgement patterns to those of the translator cluster members. The three translators in PCC2 and PCC4 also share similar judgement patterns with their

interpreter cluster mates. Among the interpreter examiners themselves, despite their different teaching backgrounds and script usage preferences, they also stay in the same clusters, such as PCC3 and PCC5.

Table 4-9.a Paired comparison clusters and the examiners' backgrounds

Paired comparison clusters – 🖋: translator 🎧: interpreter 🗣: SI teaching 📄: use script															
	1			**2**			**3**			**4**			**5**		
R1	🖋		📄												
R30	🖋		📄												
R28	🎧		📄												
R16	🎧	🗣	📄												
R15	🖋		📄												
R27	🖋		📄												
R3	🖋														
R7	🖋		📄												
R11	🖋		📄												
R29	🖋		📄												
R4				🎧	🗣										
R21				🎧	🗣	📄									
R8				🎧	🗣	📄									
R13				🎧		📄									
R5				🎧	🗣										
R23				🎧	🗣	📄									
R6				🎧	🗣										
R25				🖋		📄									
R20				🎧	🗣										
R2				🖋											
R9							🎧								
R10							🎧								
R12							🎧	🗣	📄						
R17							🎧	🗣	📄						
R18										🎧	🗣	📄			
R26										🖋		📄			
R22													🎧	🗣	📄
R24													🎧		📄
R19													🎧		📄
R14													🎧	🗣	

4.4.2 OJ clusters and the examiners' backgrounds

This section looks at the examiners' background in relation to the clusters from the OJ dendrogram (Figure 4-4.b). Table 4-9.b matches the examiners' background with their cluster membership to see if there is any pattern such as the one in the PC clusters. In 4.2.3, OJ Clusters 1 and 2 were extracted from one large cluster that has achieved the threshold single measures ICC of 0.7. With this in mind, a new pattern emerges in Table 4-9.b. It appears that the translator and interpreter examiners are roughly equally mixed in OJ Clusters 1, 2 and 4, while OJ Clusters 3 and 5 are all-interpreter clusters (hereafter OJC1, OJC2, OJC3, OJC4, and OJC5).

A shift of cluster memberships occurs between the PC and OJ assessment methods. The memberships in each of the PC clusters are more homogeneous in general, whereas on three of the OJ clusters, the memberships are mixed, each containing roughly the same numbers of interpreter and non-interpreter examiners. Although for both assessment methods there are two all-interpreter clusters, the members of those clusters are also different. For example, R12-R17 is with R9-R10 in PCC3, but in OJC3 R12-R17 has different cluster mates while R9 and R10 form a 2-member OJC5. Also, R14, R22 and R24 are in an all-interpreter PCC5, but in the OJ method they joined other translators in a mixed OJC4.

This shift of membership suggests two possibilities: either the three interpreter examiners changed their judgements from the PC method to the OJ method to be more like the translator examiners' judgement patterns, or the translator examiners changed their judgement to be more like the three interpreter examiners. This change of cluster membership may imply a change of assessment behaviour between the three assessment methods, which is generally in line with the observations in 4.1 and 4.2.

Table 4-9.b Overall judgement clusters and the examiners' backgrounds

Overall judgement clusters – 🖌: translator 🎧: interpreter 🗣: SI teaching 📄: use script															
	1			**2**			**3**			**4**			**5**		
R16	🎧	🗣	📄												
R30	🖌		📄												
R29	🖌		📄												
R2				🖌											
R28				🎧		📄									
R6				🎧	🗣										
R7				🖌		📄									
R1				🖌		📄									
R3				🖌											
R25				🖌		📄									
R13				🎧		📄									
R23				🎧	🗣	📄									
R5				🎧	🗣										
R20				🎧	🗣										
R12							🎧	🗣	📄						
R19							🎧		📄						
R17							🎧	🗣	📄						
R21							🎧	🗣	📄						
R4							🎧	🗣							
R8							🎧	🗣	📄						
R22										🎧	🗣	📄			
R24										🎧		📄			
R14										🎧	🗣				
R18										🎧	🗣	📄			
R26										🖌		📄			
R15										🖌		📄			
R27										🖌		📄			
R11										🖌		📄			
R9													🎧		
R10													🎧		

4.4.3 OM clusters and the examiners' backgrounds

Table 4-9.c shows the OM clusters (hereafter OMC1, OMC2, OMC3, OMC4, and OMC5) and the examiners' backgrounds. OM exhibits similar cluster characteristics to those in the other two assessment methods: the OM method has a mixed-membership Cluster 1 like that in the OJ method, and, like those in the PC method, has clusters that are made up of more examiners' with homogeneous backgrounds.

The cluster memberships have also shifted. Take OMC2 and PCC1 for example, both are translator-dominant clusters with one or two interpreters. Although all the five translators in OMC2 are also in PCC1, the interpreters in these two clusters are different: OMC2 has R18 while PCC1 has R16 and R28. Another interesting example is OMC3, which is a 5-member all-interpreter cluster, and these five interpreters belonged to four different clusters on the PC method (Clusters 1, 2, 3, 5) as well as in the OJ method (Clusters 1, 2, 3, 4).

Again, these are supporting evidence to show that the judgement pattern of examiners, and perhaps assessment behaviours, may change when the method of assessment changes.

* * *

From the findings in this section, in short, we found that the professional background of examiners is a rough but not a reliable guide to consistency of judgement. Translator and interpreter examiners may share similar as well as different judgement patters. Nevertheless, we have also grouped the thirty examiners into five clusters. The judgement pattern in each cluster is internally consistent regardless of the examiners' backgrounds within each cluster (4.3). For analysing and comparing the thirty examiners' perceptions of assessment criteria, it may be easier to work with a five-cluster framework than to analyse the thirty examiners individually.

Table 4-9.c Overall mark clusters and the examiners' backgrounds

Overall mark clusters – 🖋: translator 🎧: interpreter 🗣: SI teaching 📄: use script										
	1			**2**			**3**			**4**			**5**		
R8	🎧	🗣	📄												
R17	🎧	🗣	📄												
R5	🎧	🗣													
R26	🖋		📄												
R21	🎧	🗣	📄												
R23	🎧	🗣	📄												
R19	🎧		📄												
R25	🖋		📄												
R3	🖋														
R13	🎧		📄												
R27	🖋		📄												
R6	🎧	🗣													
R11	🖋		📄												
R22	🎧	🗣	📄												
R7				🖋		📄									
R30				🖋		📄									
R15				🖋		📄									
R29				🖋		📄									
R1				🖋		📄									
R18				🎧	🗣	📄									
R12							🎧	🗣	📄						
R24							🎧		📄						
R16							🎧	🗣	📄						
R14							🎧	🗣							
R20							🎧	🗣							
R9										🎧					
R10										🎧					
R4													🎧	🗣	
R28													🎧		📄
R2													🖋		

4.5 Identify a framework for analysing qualitative data

In 4.4, we have found that the internally-consistent clusters are a useful common framework for further analysis in this study, but which cluster, or assessment method, is more suitable to use? This research study aims to investigate how the examiners form their opinions and make their initial judgements on student interpreters' performances, before entering the panel discussions with other examiners (1.3). The examiners' initial judgements took place during the paired comparisons (3.2.4). To facilitate the analysis, therefore, the PC clusters are selected as the analytical framework for this study.

This section presents and analyses in more details the results of the paired comparisons, i.e. PC Thurstone's scales (4.5.1) and the student winners (4.5.2), as the framework for analysing the examiners' use of assessment criteria in this study (also see 3.2.5).

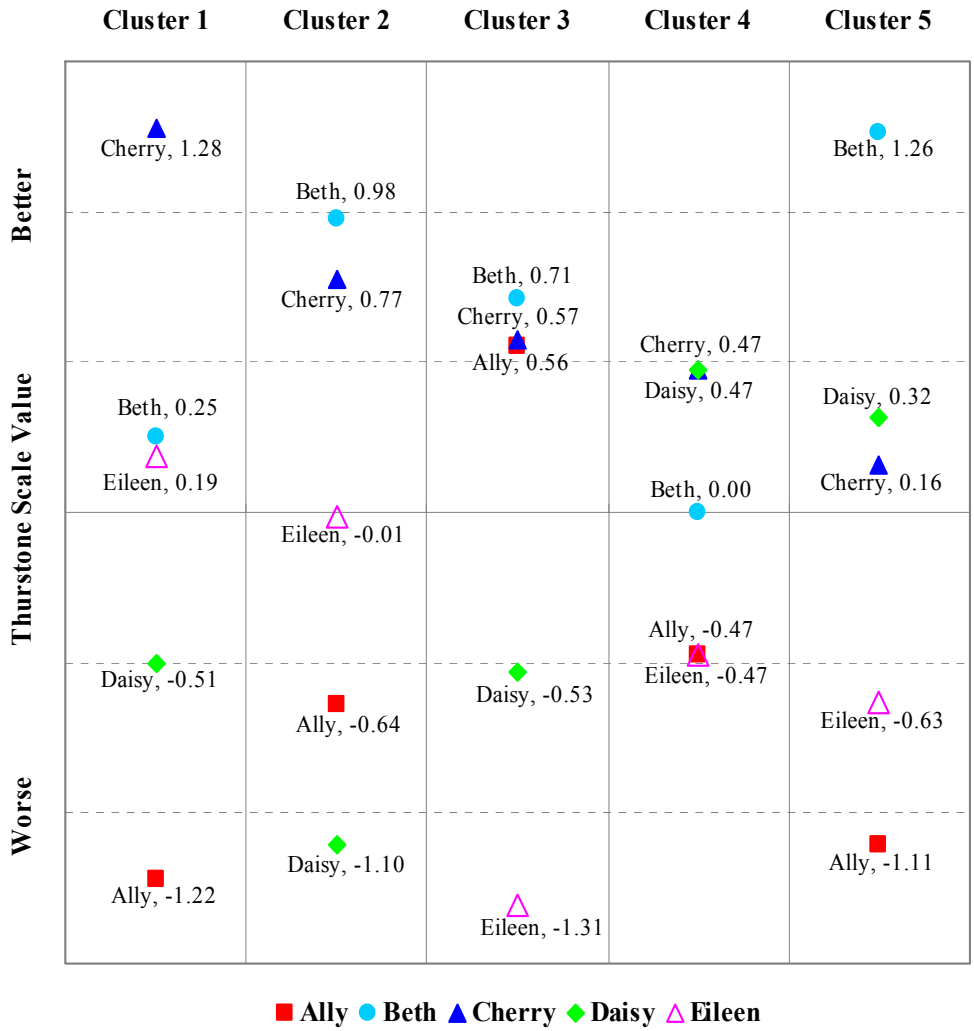
4.5.1 PC Thurstone's scales as the super examiners

As demonstrated in 4.1.2.b, Thurstone's scales (or T scales) can show the aggregated judgement of a group of examiners on a scale. In order to obtain a better picture of the examiners' judgement patterns in the PC clusters, the T scales were generated from the PC ranking points that correspond to the cluster memberships, showing the aggregated judgement of the examiners in the PC cluster.

It is expected that each cluster's T scale will show different judgement patterns because the examiners were allocated to clusters in the cluster analysis (4.3.1) based on the consistency of their judgements patterns with the other members of their cluster, and inconsistency with different judgement patterns of the other clusters.

As there are five PC clusters, there are in total five T scales (hereafter PC T scales). The PC T scales, presented in a group of five, illustrate five cluster judgement patterns, as shown in Figures 4-5 below. The PC T scales are read vertically this time: the higher the position of the student on the scale, the better the performance.

Figure 4-5 PC cluster Thurstone scales



As shown in Figure 4-5, the ranking order and the positions of the five students are all different on the T scales. This is, of course, due to the results of the cluster analysis; the examiners were sorted into different groups based on their different student rankings. Here we can use Alpha ICC again to confirm this. As discussed in 4.2, Cronbach's alpha ICC is an ANOVA-type statistical method to measure scale reliability, but the result

were expected to be *inconsistent* here i.e. having a low single-measures ICC score[20]. Using the scale values of the five PC T scales for the calculation, the single-measures ICC is low at only 0.392. This result statistically confirms that the five PC T scales are significantly inconsistent from one another.

Since each T scale represents the aggregate judgement of a group of examiners, they can be regarded as the judgements of five *super examiners*. For example, in the PC assessment method, there are five super examiners PCC1, PCC2, PCC3, PCC4, and PCC5; each of them has a judgement pattern that is different from the others, i.e. individually the five PC super examiners judge the student interpreters differently. It will be easier, therefore, to use five super examiners as a framework for analysis than to use thirty individual examiners.

4.5.2 PC ranking points according to cluster membership

In 4.1, we have reported the ranking points of the thirty examiners (Table 4-1). Here, the PC ranking points are extracted and re-arranged according to the PC cluster membership as shown in Table 4-10.

The main differences in the PC super examiners' judgements are clearly shown in Table 4-10 with bold-faced ranking points to contrast. For example, in PCC1, the major split of judgement is between Beth and Eileen, and in PCC2, the split is between Beth and Cherry; whereas in PCC3, the judgements split evenly between Ally, Beth, and Cherry. PCC4 consists of only two examiners who judged differently in two pairs – Cherry-Daisy and Eileen-Ally, and PCC5's main difference of judgement lies between Cherry and Daisy.

[20]In the cluster analysis, the single-measure ICC was set at 0.7 and above, and its consistency level refers to the *individual examiners* within each cluster. Here, the single-measure ICC estimates the consistency of the five PC *super examiners* or *clusters*, which are represented in the form of PC T scales.

Table 4-10 Paired comparison winners according to cluster membership

Cluster examiners		PC ranking points				
		Ally	**Beth**	**Cherry**	**Daisy**	**Eileen**
PCC1	**R1**	1	3	5	2	**4**
	R30	1	3	5	2	**4**
	R28	2	3	5	1	**4**
	R16	1	2	5	3	**4**
	R15	1	**4**	5	2	3
	R27	1	**4**	5	2	3
	R3	1	**4**	5	2	3
	R7	1	**4**	5	2	3
	R11	1	**4**	5	3	2
	R29	1	3	5	3	3
PCC2	**R4**	3	4	**5**	1	2
	R21	3	4	**5**	1	2
	R8	2	4	**5**	1	3
	R13	2	4	**5**	1	3
	R5	2	**5**	4	1	3
	R23	2	**5**	4	1	3
	R6	1	**5**	4	2	3
	R25	1	**5**	4	2	3
	R20	1	**5**	3	1	3
	R2	2	4	3	1	5
PCC3	**R9**	5	**5**	3	3	1
	R10	4	**5**	3	2	1
	R12	3	4	**5**	2	1
	R17	4	3	**5**	2	1
PCC4	**R18**	**1**	3	4	**5**	2
	R26	2	3	**5**	4	**1**
PCC5	**R22**	1	5	**4**	3	2
	R24	1	5	3	**4**	2
	R19	2	5	**4**	3	1
	R14	1	5	2	**4**	3

Some interesting within-examiner inconsistencies are also noticed in Table 4-10, which are indicated by underlined ranking points. When referring to the actual winners of the paired comparisons (see Appendix D), we see that Examiner R29 could not successfully separate the students in two pairs: Daisy-Eileen and Eileen-Beth, but

interestingly, R29 could judge Beth to be better than Daisy. By logic, therefore, Beth should have been better than Eileen when Daisy and Eileen were considered a tie.

Two other cases of internal conflict of judgement were also observed – in PCC3 R9's judgements of Beth-Cherry-Daisy, and in PCC2 R20's judgements on Ally-Cherry-Daisy. R9's judgements were Beth>Cherry, i.e. Beth was better than Cherry, and Cherry>Daisy. Therefore, by logic Beth>Daisy should be inferred. However, R9 judged Daisy>Beth in another comparison. This was in conflict with R9's own previous judgements by logic. As for R20, when comparing Cherry and Daisy, the examiner judged Cherry>Daisy. However, in the other two paired comparisons, R20 judged Daisy>Ally, and Ally>Cherry, which do not add up by logic. These are typical examples of within-examiner inconsistency.

The analysis of the cluster-based PC results will guide part of the qualitative analysis and discussion on the examiners' use of assessment criteria later in Chapter 5.

4.6 Analytical framework for examiners' use of assessment criteria

In this chapter, the results showed that the consistency between assessment methods of the thirty examiners as a group is excellent. However, obvious between-examiner inconsistencies were also observed. A simple predictor of the consistency of group membership, the examiners' backgrounds, was used to for analysis in 4.1 and 4.2; the background factors used were whether the examiners were interpreters or translators, had teaching experience of simultaneous interpreting or not, and whether or not they preferred to use the script. It is interesting to see that non-interpreter examiners, who mostly rely on using scripts, appear to be more consistent than the interpreter examiners. However, it is difficult to say with confidence that interpreter examiners who use scripts and have teaching experience are more consistent than those who do not use scripts and do not teach. There seems to be more than simply their external backgrounds affecting the examiners' consistency level.

In 4.3, new group memberships, i.e. the clusters, were identified for each assessment method. Within each cluster, examiners share a more similar judgement pattern regardless of their external background. After cross-examining the cluster membership with the examiners' background (4.4), some interesting patterns were noted. Some clusters were more homogeneous, and consisted of mostly interpreters or mostly non-interpreters, and some were equally mixed. In other words, interpreter examiners may judge similarly or differently among themselves, and non-interpreter may have similar judgement patterns to the interpreter examiners. The cluster judgement patterns, therefore, may not entirely depend on the factor of examiner backgrounds.

During the data exploration, a question arose as to what may be the minimum or optimum number of examiners required to achieve a consistent judgement result

(4.2.1.b). Since the clusters of examiners were extracted based on the similarity of their judgement pattern, the number of examiners in the clusters may be a good way to answer the question. However, the number of examiners in the clusters ranges between two and fourteen (Tables 4-7) and so does not provide a practical answer to the question. The question was further complicated by the findings in 4.4, which showed that there is no clear relation between the examiners' judgement patterns and their backgrounds.

This leads to revise the question as: how do we select the limited number of examiners for the examination panel that can achieve consistent judgement results? One way to do this is, given the fact that "practicality" limits the numbers of examiners, to select a maximum of three examiners, for instance, from the most dissimilar clusters, i.e. to obtain the widest spread of views on the student interpreters' performances, and then to average out the marks of the three examiners. This approach is taken from a perspective of achieving high test validity as the results are based on a wider base of professional judgements (see 2.3.2.b). Another approach is to select the three examiners from the same cluster, or similar clusters, to ensure the consistency of marking, i.e. taking a high-reliability approach.

This of course leads to another question: which approach should be taken? The answer lies in the balanced view of test validity and reliability (2.3.4). It depends on the purpose of the interpreting examination, and needs to be decided based on the professional judgement of the test designers and examiners (2.3.2.b). This brings us back to the concerns of subjectivity issue in the examiners' judgements, which need to be made by using some assessment criteria to stay as objective as possible (2.4.3). Therefore, the question of what the criterion or criteria may be arises, which is related to the next research question to be answered in this study.

In 4.5.1, it was established that the examiners as a group in PC cluster, or the PC super examiners, all behave differently in terms of student ranking judgement. As the

super examiners are made up of the individual examiners, it would be interesting to see whether they use the assessment criteria differently or similarly. In other words, understanding how the examiners exercise their judgements may also help answer the question of how a reliable panel of examiners can be selected, or how an effective examination procedure may be designed to achieve more reliable examination results.

Therefore, the logical next step would be to look at the data for the examiners, i.e. their views on the student interpreters' performances, to understand their use of assessment criteria better. For this purpose, apart from the coding method (3.2.5), an analytical framework is needed for investigating the internal judgement patterns of the thirty subject examiners using their verbal comments. Compared with thirty examiners, five super examiners is a more manageable framework to analyse the examiners' use of assessment criteria. For example, it was found that within each of the PC clusters, there were still different opinions on the performances of certain students (4.5.2) even when the rankings are statistically consistent. These findings provide this study with a way to look into how the examiners judged the five students, and they will serve as a framework for analysing how the examiners' use the assessment criteria, which will be carried out in Chapter 5 next.

Given the above findings, the three research questions set out at the beginning of this chapter can be answered as follows: (1) five different yet internally-consistent judgement patterns of the examiners have been identified for each assessment method (2) non-interpreter examiners are slightly more consistent than the interpreter examiners in assessing the five student interpreters, especially when they use the speech scripts, and then (3) the five PC clusters of examiners, i.e. the five PC super examiners, may be used as a common framework for the analysis and discussion of the qualitative data on the examiners' use of assessment criteria.

CHAPTER 5

Assessment Criteria

for Simultaneous Interpreting Examinations

5.0 Introduction

The contribution of this chapter to the overall aim of the present study will be to fulfil the third research objective, namely to elicit and understand the important assessment criteria used by the examiners (1.3). Previously in Chapter 4, different ranking patterns were identified among the thirty examiners, and the patterns were sorted into five super-examiner types as an analytical framework for discussing the qualitative data. In fulfilling the research objective mentioned above, this chapter will identify (1) what assessment criteria were actually used by the examiners, and then (2) ascertain the relationships between the assessment criteria and the judgement results. For example, were the ranking patterns the results of differences in examiners' assessment criteria, or did the examiners use similar criteria, but on that basis rank students differently?

In this study, the examiners first listened to the student interpreters' performances and the paired comparison was the first point at which the examiners judged the performances. Therefore, the primary dataset for analysis is the examiners' paired comparison comments. The comments of the individual examiner within each PC cluster were examined to explore the assessment criteria they used, and how they judged the students.

The method of coding and analysing the qualitative data will be explained in 5.1. Then, the concepts and categories that emerged from the qualitative data, i.e. the assessment criteria, are presented in 5.2, 5.3, 5.4, 5.6, and 5.6. Toward the latter part of the chapter in 5.7, the salient assessment criteria used during the examiners' decision-making process are identified, followed by a summary discussion in 5.8 to answer the research question mentioned above.

5.1 Coding of the interview data

The examiners' paired comparison comments were transcribed for analysis. As reviewed and explained in 2.5.3.c and 3.2.5, the current study adopted the coding practice of Grounded Theory (GT) to scrutinise the unstructured transcribed data in the hope to identify the concepts of assessment criteria that the examiners used. Normally, the GT method is an iterative process for collecting and analysing qualitative data (such as interviews) until a theory is derived from or "grounded" in the data (Bryman, 2004: 401). The current study, however, collected the interview comments only once during the paired comparisons, and adapted the GT principle (Charmaz, 2006: 9) to code the interview comments by recursively analysing the qualitative data – also a key process in the Grounded Theory approach.

There are three distinctive types of coding practices in the Grounded Theory – open, axial, and selective coding (Bryman, 2004: 402). Open coding involves line-by-line examination of the text data, breaking them down, and comparing them, and from these coding processes emerge *concepts*; similar concepts then can be further grouped into *categories*. Axial coding makes connections between categories to identify any pattern, whereas in selective coding, core categories are selected and related to one another to explain their relationships.

In this section, the initial open coding process will be illustrated and followed by explanations of how the identified concepts are grouped into categories for analysis and discussion.

5.1.1 Initial open coding

Table 5-1 Example of initial open coding – Cherry and Ally

Examiner's Comment	Coded concepts
English translation: 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. **Original in Chinese:** C 整體還是比 A 要好得多了。她的聲音比較甜，很穩定，速度很平均，不會忽快忽慢，也很少有非常長的停頓。那比較不會有沒有意義的、空洞的口頭禪、字眼在那邊填補。那我的印象中前面的三分之二他做得相當地不錯，那到後頭可能一些重大的訊息數字他也知道自己聽漏了吧，35%到 40%的市場佔有率，他說成了 45%，而且不知道是不是在那裡努力地回想這個數字，所以那一句相對於他其他的句子，那一句中文就顯得很不順。	**PD:** sweet voice, steady, stable pace, pauses, fillers, fluency **EB:** examiner impression **FAI:** listening comprehension, **FC:** omissions, message weighting, numbers, terminology, **EB:** examiner speculation PD: Presentation & Delivery EB: Examiner Behaviour FAI: Foundation Ability for Interpreting FC: Fidelity & Completeness

The interview transcripts of the thirty individual PC examiners were examined line-by-line. Table 5-1 shows an example of the initial coding process where an examiner was comparing students Ally and Cherry. When a distinctive idea or concept was identified (underlined in the table), the conceptual property was coded by using a key word or phrase. The idea or concept was 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). Most of the examiners' comments were

made in Chinese. For easier reading and reporting, the codes are in English and an English translation of the comments is included here with the original comment in Chinese. The conceptual properties and codes are annexed to the comments in the Coded concepts column as shown in Table 5-1.

From the comments in Table 5-1, for example, the examiner noted how the messages were delivered (voice, steady, pauses, fluency), and the completeness of the messages (missed some important numbers: omission). Some of the examiner's assessment behaviour was also noticed, such as speculating on why the interpreter made some mistake (toward the end she was probably also aware that she did not hear the numbers) and surmising on the reasons why the fluency of her delivery was affected (this might be because that she was busy remembering the numbers). The "my impression" comment also indicated that the examiner acquired a general impression of the students' performances and used that impression to compare the students and made judgements. Concepts like these were underlined and coded so that they could be collated with those that were identified from the other examiners' comments.

5.1.2 Collating and sorting concepts into categories

After the line-by-line coding of the thirty examiners' comments in the five paired comparison clusters, the coded concepts were then compared and collated with one another; similar concepts were grouped into categories. As the concepts in the same category are similar in nature and describe one specific aspect of the student interpreters' performance, the conceptual category can be regarded as a multi-dimensional assessment criterion (see reviews in 2.1, p.18-29, regarding the multi-dimensional perspectives on interpreting quality assurance) which contains finer aspects of judgement. The number of categories, i.e. assessment criteria, is much

smaller than the number of concepts, which makes it easier to analyse their relationships in the assessment process.

Taking the concepts in Tables 5-1 for example, some are more related to describing the way the student interpreters deliver the messages, such as voice, pace, fillers, calm, confident and hesitations. These concepts can be grouped and referred to as the "presentation and delivery" category. Therefore, an assessment criterion is identified as *Presentation and Delivery,* which consists of different conceptual properties (i.e. different aspects) that describe how the student interpreters are being assessed in terms of the presentation and delivery of the interpreted messages.

There are other concepts that comment more on the fidelity and completeness of the delivered messages, such as translating the numbers wrongly, translation errors, omissions and confusion. This type of concept can be sorted into another category and called *Fidelity and Completeness* category, i.e. another assessment criterion, which also contains different aspects for judgement when using this criterion.

There are, of course, more than two conceptual categories. All the concepts identified from the thirty examiners' comments were collated and sorted into different categories by using the processes explained above; ultimately, six conceptual categories emerged. They are: *Presentation and Delivery, Fidelity and Completeness, Audience Point of View, Interpreting Skills and Strategies, Foundation Abilities for Interpreting* and *Examiner Behaviour*.

The first five categories can be regarded as the assessment criteria, on which the examiners based their judgements. However, the last category – Examiner Behaviour – is not an assessment criterion. The concepts in this category can be regarded as *factors* that may influence the judgement process, which will be presented and analysed in Chapter 6. This chapter presents the five assessment criteria that emerged and discusses their conceptual properties to understand how the examiners used them.

5.2 Presentation and Delivery

Table 5-2 Conceptual properties of Presentation and Delivery

Super Examiners	Concepts of the Presentation and Delivery Category
PCC1	Accent, coherence, transitions, comfortable to listen to, control of breathing, sigh (microphone usage), tones, pleasant, sweet, concise, convincing, not hesitant, with deliberate steps, disconnected sentences, focused/unfocused, idiomatic, unidiomatic usage, word choice, word order, regional usage, diction, pace, rhythm, tempo, fluency, slow, fast, fluent, mumbling, pauses, rendition, polished, literal rendition, simple expression is better, voice quality, nervousness, tense, steady, high pitch, sound tired, natural voice quality
PCC2	Accent, poor breathing control, microphone usage (too far away), calm, panic, excessive fillers, not confident, nervous, sighs even before the interpretation began=lose credibility, diction, word choice – regional usage, lack of vocabulary, wordiness, clear and logical, more organised, transitions, vague delivery, concise rendition, not polished enough, expressions, fragmented Chinese, even, steady pace, not rushed, better delivery, get stuck, not fluent, Chinese not good enough, warm-up longer, dangling sentences, jerky delivery=incomplete sentences
PCC3	phonology, clear voice, accent, intonation, short of breath, panic, nervous, confident, not convincing, regional choice of words/expressions, cliché, wordiness, need to improve business terms, tempo, rhythm, blank delivery, pauses, steady, slower pace, dragging, more comfortable to listen, concise delivery in a positive way, keep to the speaker's delivery style
PCC4	voice projection, immature tone, annoying intonation, calm, nervous, confidence, convincing, steady, uneven pace, hesitation, fluent, stammer, rhythm, fillers, vague delivery, sloppy, not concise enough, word choice difficult to understand, unclear, incorrect expression, delivery too poor to understand, incomplete sentences, translates half,
PCC5	voice, pace, pronunciation, booth manner, nervous, unsteady, convincing tone, word choice, usage of terms, excessive fillers, clear, vague delivery, understandable, difficult to understand, serious interference in delivery, steady performance

Table 5-2 shows the concepts in the Presentation and Delivery category. In this category, the examiners made various references to the way student interpreters presented and delivered the messages (hereafter the Presentation and Delivery criterion). As simultaneous interpreting does not require interpreters to speak directly to the audience, the term *Presentation* here refers to the interpreters' vocal presentation rather than their public speaking presentation; the term *Delivery* refers to any other concepts that come hand-in-hand with the vocal presentation, which predominantly relate to the interpreters' language usage, or the texture of the target language output.

5.2.1 Three dimensions of the Presentation and Delivery criterion

Table 5-2.a Dimensions of Presentation and Delivery

Criterion	Dimensions	Conceptual properties
Presentation and Delivery	**acoustic**	Accent, comfortable to listen to, control of breathing, sigh (microphone usage), tones, pleasant, sweet, convincing, not hesitant, pace, rhythm, tempo, fluency, slow, fast, fluent, mumbling, pauses, voice quality, nervousness, tense, steady, high pitch, sound tired, natural voice quality, phonology, clear voice, intonation, short of breath, panic, nervous, confident, not convincing, more comfortable to listen, booth manner,
	word/phrase usage	idiomatic, unidiomatic usage, word choice, word order, regional usage, diction, rendition, polished, diction, word choice – regional usage, lack of vocabulary, wordiness, expressions, Chinese not good enough, regional choice of words/expressions, cliché, wordiness, need to improve business terms,
	Flow of information	coherence, transitions, concise rendition, with deliberate steps, disconnected sentences, focused/ unfocused, literal rendition, simple expression is better, excessive fillers, sighs even before the interpretation began=lose credibility, clear and logical, more organised, transitions, vague delivery, not polished enough, fragmented Chinese, steady pace, not rushed, better delivery, even, get stuck, not fluent, warm-up longer, dangling sentences, jerky delivery= incomplete sentences, tempo, rhythm, blank delivery, pauses, slower pace, dragging, keep to the speaker's delivery style, serious interference in delivery

Through further comparisons and a focused coding of the concepts within the concepts in Table 5-2, three dimensions of the Presentation and Delivery criterion emerged: the acoustic dimension, the word/phrase usage dimension, and the flow of information at sentence level. In other words, when the examiners form their judgements on this dimension, they pay attention to at least these three different aspects of the student interpreters' performances. With some variations in the conceptual properties, all five clusters of examiners' comments reflected the three dimensions of the interpreters' presentation and delivery. Table 5-2.a illustrates how the conceptual properties were sorted into the three dimensions of Presentation and Delivery.

The acoustic dimension

The concepts in this dimension are perhaps the first that examiners notice when they listen to an interpreter's performance. On the acoustic dimension, an examiner hears the pitch and tones of the interpreters' voice and forms a first impression, such as sweet or "girly" voices. This acoustic impression further extends to various vocal expressions, such as hesitation, nervousness, tiredness, and even sighing in frustration.

The word/phrase usage dimension

On the word/phrase usage dimension, an examiner notes the idiomatic usage of the target language, such as terms, idiomatic expressions, regional usage, and the rendition of the message. Where the delivery of the message is unpolished or too literal, the judgements tend to be negative. For example, some examiners made comments like this: "Her accent is from Mainland, if she is interpreting in Taiwan, the audience may not be used to it[21]", and "Her Chinese [interpretation] usage sounds very natural, [which] was not influenced by English[22]".

Flow of information at sentence dimension

On the third dimension, examiners focus on the flow of information within and between sentences. Generally speaking, concise and simple expressions delivered at a steady and fluent pace are preferred, such as the flowing sample comment in Table5-1:

> 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.

When reviewing the linguistic standards for conference interpreting, Hartley points out

[21] Source text in Chinese: 她口音聽起來是大陸的，在台灣翻的話觀眾聽起來可能不習慣。
[22] Source text in Chinese: 她的中文用語聽起來很自然，沒有受英文影響。

that "making a sequence of sentences operational" (both cohesive and coherent) is important to simultaneous interpreting (Hartley et al., 2004), a view consistent with the findings here. Examiners attend to how well student interpreters make sense groups or messages clear within and between sentences, such as by using cohesive linguistic devices, transitions, and the logical progression of the way the messages are delivered, or, as one of the PCC1 examiners' commented "conveying with a measured tread" (按部就班的傳遞).

If there is a combination of positive comments across the above three dimensions of presentation and delivery the interpretation is more pleasant, convincing and comfortable to listen to, and the judgement will be more positive, and vice versa.

5.2.2 Variations in applying the Presentation and Delivery criterion

The five clusters of examiners shared the three dimensions of the Presentation and Delivery criterion. Nonetheless, in examining the interview data, it was noticed that the PCC1, PCC3 and PCC5 examiners' comments appeared to be more evenly distributed between the three dimensions; whereas the PCC2 and PCC4 examiners seemed to be more emphatic when commenting on the third dimension with more frequent references[23] to the importance of being fluent in delivering the interpretation, and to how difficult or easy it was to understand the delivery. They emphasised that the ideal interpretation delivery should be clear, logical and organised, the transitions smooth and the flow of messages fluent at a steady and even pace. A concise rendition is preferred, not wordiness.

[23] The frequency of the references to concepts does not show in the table. The concepts in the table only represent various observations on the students' performances. One concept may appear in the table once, but may be mentioned by the examiners on many occasions. The statement "more frequent references" here is based on the researcher's observation when examining the interview data.

The examiners also made various links between delivery and other criteria. Concepts such as "vague delivery" and "fragmented Chinese" are examples of links to fidelity and completeness. Deliveries that are "jerky" or have long pauses imply that the sentences are incomplete, and the messages may also be incomplete because they are too poorly delivered to be understood, which echoes Schjoldager's argument (2.4.2.c) regarding the importance of the delivery criterion. Some examiners also commented that the interpreter should try to follow the speaker's style of delivery, otherwise, the intended message may not be fully conveyed. It seems that these links between the delivery and presentation criterion and the fidelity criterion mean the two criteria cannot be completely separated. They need to be considered in relation to each other.

In addition, one student interpreter made a loud sigh at the beginning. This sigh prompted a comment that an audience would lose confidence in this interpreter, i.e. a link between the Presentation & Delivery criterion and the Audience Point of View criterion. The link between personality and interpreting delivery was also mentioned, such as getting into panic easily and having an "immature tone." This link is related to the criterion of Foundation Abilities for Interpreting (5.6) about the aptitude for being an interpreter.

Given the above variations of focuses, such as PCC2 and PCC4 on dimension three, and the links in the criterion's conceptual properties to the other criteria, it appears that different super-examiners formed their judgements in slightly different ways from one perspective to another. This is in line with the multi-dimensional perspective on the quality assurance of interpreting that was reviewed in 2.1.2, which may be an indicator for further exploration of the reasons for the fluctuations of the individual examiners' judgements observed in Chapter 4.

5.3 Fidelity and Completeness

Table 5-3 Conceptual properties of Fidelity and Completeness

Super Examiners	Concepts of the Fidelity and Completeness Category
PCC1	content accuracy, numbers/figures, terms, significant point, jest, keep to the theme, meaning errors, mistranslation, slip of tongue, omission, incomplete sentences, fragment, over-translation, more information, fabrication, intention of the speaker, greetings, introduce company/product; context, make sense, context of numbers, inconsistent context, message weightings, different levels of fidelity, numbers important for business interpreting, consistency, stable, steady, focused/unfocused
PCC2	content accuracy, accuracy rate, abandoned messages, incorrect messages, missed key words, numbers, complete, level of fidelity=at least something rendered, minor/major errors, different consideration in different exams, information in the beginning not important, missed a lot but luckily not important messages, incomplete but got the main idea, making up stories, completely the opposite meaning, package four levels of messages, basic mistakes, message rendered ok but meaning a bit different from the source, numbers important to business meetings, gave more information, no context, unclear context, clear/unclear message, vague but no obvious mistake, frequency of mistakes, got the point but rendition not polished enough, combine sentences=change meanings, too many small mistakes=cannot keep listening, message too vague, jerky delivery=incomplete sentences, function of the speech act: greeting vs. information, peer-referenced criterion (experienced interpreter, market), preparation could remedy lack of company background knowledge,
PCC3	accuracy, missed numbers, missed the greetings, names and titles, wrong company product, key words, comprehensive, complete/incomplete sentences, fragmented content, message closer to the speaker, syntax, semantic, deeper layers of meaning, no focus, slip of tongue, messy logic, error in logic, numbers jumbled up, no context, fabrication, add stuff, different criteria weightings, accuracy cover rush delivery,
PCC4	lost messages, missed more things, mistranslation, not serious mistake, serious mistake, incorrect numbers, incomplete sentences, , almost all wrong, get the gist, talking nonsense, logical problem, fabrications, context did not add up, only translate half, missed a lot, prefer omissions than errors, more complete more errors, lost a lot of messages but less errors
PCC5	incorrect numbers/translation, completely wrong, missed something, more complete, the opposite meaning, mistakes, missed the theme, did not get the meaning correctly, message no focus, incoherent, coherent logic between sentences, all very vague, vagueness is worse than omission

Table 5-3 shows the concepts that are related to faithfulness and accuracy of the messages conveyed by the interpreters in relation to the source speech, grouped according to their mention by the five super-examiner clusters. These concepts are at the core of quality when assessing interpreters' performances (2.1, 2.2). The concepts in this category refer to similar things, but interpreter trainers, professionals and users of interpreting service appear to use the terms in various ways when describing this assessment criterion (Hartley et al., 2004). For example, among other things, the term *accuracy* seems to be mentioned more often than the term *faithfulness* due to the "moral overtones" of the latter (ibid: 10-11). Other terms are also used by different researchers, such as *sense consistency* (Kurz, 2001; Moser, 1995) or *loyal/disloyal* (Schjoldager, 1995). Gile used the term *fidelity* when discussing the concept and its principles in both interpretation and translation (1995: 49-74), which encompasses a wider spectrum of considerations on the quality of the process when rendering messages from the source language into the target language.

Considering the above, therefore, this study will adopt Gile's choice of the term, and use Fidelity (the quality of faithfulness) and Completeness (of sentence structure and messages) when referring to the concepts in this category, or assessment criterion.

5.3.1 Three main properties of the Fidelity and Completeness criterion

Through recursive comparisons of the concepts collated in Table 5-3, it was found that the examiners' attention mainly focused on three main conceptual properties: content accuracy, speaker intention, and contextual consistency, albeit with some variations of focus. Table 5-3.a below shows how the conceptual properties in the Fidelity and Completeness are sorted into the three main dimensions of this assessment criterion.

Table 5-3.a Dimensions of Fidelity and Completeness

Criterion	Dimensions	Conceptual properties
Fidelity and Completeness	**Content accuracy**	content accuracy, numbers/figures, terms, meaning errors, mistranslation, slip of tongue, omission, incomplete sentences, fragment, over-translation, more information, fabrication, consistency, stable, steady, focused/unfocused, incomplete but got the main idea, accuracy rate, abandoned messages, incorrect messages, missed key words, numbers, complete, level of fidelity=at least something rendered, minor/major errors, making up stories, completely the opposite meaning, basic mistakes, message rendered ok but meaning a bit different from the source, vague but no obvious mistake, frequency of mistakes, got the point but rendition not polished enough, combine sentences=change meanings, too many small mistakes=cannot keep listening, message too vague, jerky delivery=incomplete sentences,
	Speaker intention	significant point, jest, keep to the theme, intention of the speaker, greetings, introduce company/product, function of the speech act: greeting vs. information,
	Contextual consistency	make sense, context, inconsistent context, no context, unclear context, clear/unclear message, different levels of fidelity, message weightings, context of numbers, numbers important for business interpreting, different consideration in different exams, information in the beginning not important, missed a lot but luckily not important messages, package four levels of messages, peer-referenced criterion (experienced interpreter, market),

Content accuracy

In this conceptual property, an examiner points out the obvious mistakes that the interpreters make, such as numbers, terms, distortions, meaning errors, omissions, and over-translations. Among the various mistakes, examiners hold the view that fabrications or "making up stories" are the most serious. Comment 1 below illustrates this view point.

Comment 1 (translation)[24]**:** To me, D gave me a worse impression at the beginning part. She did not get the main ideas from the speaker. However, she still

[24]**Comment 1** in source text Chinese: 我想 D 對我來說，給我印象比較差的就是他在前面的部份，因為前面一開始的部份他那個講者想要講的重點他沒有抓住，可是還是想辦法說一些東西，而且說得頭頭是道，他那種 assured 的態度讓我覺得很可怕。就是八竿子跟講者講的東西完全沒有關係，可是他還是講得好像非常地自然，我覺得這是，做聽眾來講可能沒有感覺，我是我們做教書的就會覺得這是最可怕的口譯員。

> tried to say something and made up a convincing story. What worries me most is her assured attitude when she could keep saying things that were totally irrelevant to what the speaker said. The audience might not be aware of the difference, but from the perspective of an interpreter teacher, I feel that this type of interpreter is the most horrible one.

Usually the examiners would attribute this kind of mistake to the interpreters' poor listening comprehension and to their lack of good simultaneous interpreting skills and strategies, i.e. links to other assessment criteria.

Speaker intention

The second main conceptual property in this category is about being consistent with the speaker's intention. This concept also appears in Hartley's peer-assessment grid under the item of "rhetorical force" which includes intention and emotion, the former being the speaker's speech act and the latter the attitude of the speaker (2004: 23).

In this study, however, the speaker's emotion was not mentioned by the examiners. This might be because the speech used in the study was the first three minutes of a formal speech in a business meeting and the speaker used a neutral and non-emotional tone. Therefore, the subject examiners only focused on the speaker's speech-act and the factual information, such as giving greetings, purpose of the trip, and company introduction. Examiners consider these as being faithful to the speaker. Hence a part of the Fidelity and Completeness criterion.

Contextual consistency

Contextual consistency is the third area to which the examiners' referred. The examiners hold the view that if the messages are inconsistent with the context, they would not make sense, and the audience would be confused. This echoes the third level

of the Presentation and Delivery criterion, which emphasises the fluent and logical delivery of the messages (5.2.1).

However, there is a distinction: the Presentation and Delivery criterion pays attention to how the interpreters use linguistic devices to make the sentences operational, whereas the Fidelity and Completeness criterion looks at the contextual or factual consistency of the messages. In other words, the former relates to the form of words and the latter to the consistency and meaning of the words in relation to the source speech messages. It is not just the errors in individual messages that concerned the examiners, but also the consistent context of the interpreters' messages in relation to the source speech messages.

Comments 2 and 3 below illustrate the point. An interpreter may be able to interpret fluently and confidently with "accurate" numbers or terms, but if they are out of context and the messages are inconsistent with the source speech, the interpretation will still confuse the audience.

> **Comment 2 (translation)**[25]**:** The first interpreter A (Ally), she got the number about the international market share, but then she said 90% without a clear reference to what it was about. So the audience might be a bit confused. It was 35% to 40%, why 90% now? Interpreter B (Beth) got both numbers wrong. The two numbers were totally wrong. The third number is the turnover, the company's turnover last year. Both students didn't get the figure and both got it wrong. I mean, *it's not so much as in getting the numbers wrong, but in the key point that they got the context of the numbers wrong.** One said about spending, the other said about the profit in the US. Both were wrong.
>
> *Italics are the researcher's added emphasis.

[25] **Comment 2** in source text Chinese: A 呢前面第一個，就國際市場佔有率的那個數字是抓住了，但是中間他又說了一個 90%的這個數字呢他沒有說清楚到底是什麼，所以聽眾可能有一點糊塗，剛剛說 35%到 40%，為什麼又變成 90%。B 這個口譯員在說這兩個數字的時候呢也都說錯了，這兩個數字完全說錯了。第三個數字就是他們的 turnover 去年的 turnover，兩個都沒有說到，而且兩個都是說錯了，就是我覺得*數字沒抓住還是次要的，他們關鍵把這個數字的上下文內容是什麼說錯了*。一個說花了多少錢，第二個說在美國挣了多少錢。這兩點都說錯了。

> **Comment 3 (translation)[26]:** Usually the (interpreter's) first sentence is very important. What was her interpretation of the first sentence? […] She said something like people from the academic field […]. However, the speaker did not say anything about the academics. In this way, […] making up things with a big gap like this will cast doubts in the mind of the audience, wondering […]: was the previous interpretation wrong? Or is the current interpretation not right? Therefore, it was not handled well here.

Comment 3 also shows how an examiner, or a user of the interpretation service, may look at the contextual accuracy beyond the more obvious examples of numbers and terms. It is out of the context of a business meeting to mention academics; a reference the speaker did not make in the first place. The important thing in the mind of the examiner is that the interpretation needs to be linguistically cohesive and contextually coherent so the messages can be successfully conveyed and received across the language barriers.

5.3.2 Different levels of importance in fidelity

Just as in the Presentation and Delivery criterion (5.2.2), there are variations among the examiners in the ways they described the three different areas of conceptual properties in the Fidelity and Completeness criterion. In terms of the variations between the super examiners, all five clusters of examiners share similar conceptual properties, but the examiners' comments in PCC1, PCC2 and PCC3 are more similar to one another with a clear distinction between the three areas. PCC4 and PCC5 examiners' comments seem to be more related to the third area of contextual consistency.

[26] **Comment 3** in source text Chinese: 第一句通常很重要，第一句他翻成什麼？[...]來自學界[…]的人什麼的，可是在講者他所講的東西裡面完全沒有學界。那你這樣子，[…]掰的差得很遠，聽眾會開始懷疑自己，想說，[…] 是之前的翻譯翻錯，還是現在的翻譯翻錯。所以這個地方處理得不是很好。

Some examiners pointed out that there are "different levels of fidelity", i.e. the examiner regarded different messages as having different levels of importance in the communication process. For instance, an error caused by a slip of tongue during interpreting may be forgivable when compared with a mistake that results from misunderstanding of the source speech.

Another example is the interpretation of numbers. They may not need to be 100% exact in a general discourse, but they must be reasonably accurate so that the messages carried by the numbers are not distorted or missing. Also, it is considered to be more serious to get the numbers wrong in a business meeting than in a non-business meeting where an approximation may suffice.

Examiners also talked about errors as major ones and minor ones. A PCC2-type examiner, for example, appears to comment more frequently on content accuracy. PCC2 examiners not only looked at the number of errors, but also considered the types of mistakes, including describing them as significant or insignificant errors. Content accuracy seems to be the more dominant conceptual property for a PCC2-type examiner. Comments 4 and 5 below made by two individual examiners in PCC2 are presented to illustrate this point.

> **Comment 4 (translation)[27]:** If I were to hire an interpreter, I would pick B because B only has minor mistakes. C is good overall, but she made several major mistakes that shouldn't have happened. She made mistakes in key areas.

> **Comment 5 (translation)[28]:** I felt that C processed very quickly and her rendition was simple and concise without major omissions of information. Of course both had some errors on numbers but C's errors were more minor.

[27] **Comment 4** source text in Chinese: 如果我要請口譯員的話，我會選 B，因為 B 只有小錯誤。但是 C overall 都很好，就是有幾個重大不應該錯的地方，關鍵的東西他錯了。

[28] **Comment 5** source text in Chinese: C 我覺得他處理得非常快，然後翻出來的簡單扼要，沒有很重大的資訊的遺漏。當然兩者在數字上都有些誤差，但是 C 的誤差比較小。

These two comments explicitly distinguish messages and errors as major ones and minor ones; so, the examiners clearly have different levels of fidelity in mind regarding the messages to be interpreted.

Apart from the above distinctions, Comments 4 and 5 also reveal a discrepancy in the judgements of the students' performances. Comment 4 indicates that Cherry made some major errors but Comment 5 says that Cherry processed the information quickly and delivered it in a concise way without major omissions. The errors and omissions mentioned in the comments may or may not relate to the same messages in the interpretation, but this discrepancy shows that there may be more to explore to understand how examiners use the criteria to assess student interpreters.

In addition, some examiner behaviours were noted that might affect how a mark would be assigned using the Fidelity and Completeness criterion. For example, examiners speculated on the student interpreters' use of strategies that might have resulted in the success in or failure to conveying the messages at different fidelity levels. At times, the speculation was that the use of strategies may create future problems but that the problems were not evident in the observed performances. The factor of examiner behaviours will be discussed in more details later in Chapter 6.

5.4 Audience Point of View

Table 5-4 Conceptual properties of Audience Point of View

Super Examiners	Concepts of the Audience Point of View Category
PCC1	gain audience's confidence, warm-up fast, delivery impact, more comfortable to listen, willing to keep listening (meaning doesn't really matter), expect to be like a professional, beginner/in-training, unprofessional girlish fillers, breathing control, sigh (microphone usage), concise, consistent and coherent speech, transition, confuse audience, fragmented, idiomatic usage, regional usage to target audience, diction, pace, fast=nervous, slow=relaxed, pace, rhythm, tempo, fluency, slow, fast, fluent, mumbling, pauses, polished=easy to understand, unfocused, faithfulness to SL, introduce company product, purpose of speech, can get more messages
PCC2	less confidence if pauses are long, word choice more suitable to local audience, sound more fluent, sighed before the interpretation began=loss credibility, audience wouldn't know that it's fabrication if the delivery is good but teachers would, startle the audience by saying something out of the context, can follow the interpreter better (less errors and omissions due to overlapping), more clear and logical from the audience point of view, interpreting is service industry=can't make the audience nervous when listening, audience would feel more comfortable to listen, alternate point of view: teacher, examiner and audience, gave a sigh even before the interpretation began=loss credibility,
PCC3	feel pressure / no pressure, nervous interpreter makes nervous audience, can hear and understand more clearly, more comfortable to listen, keep the audience waiting, prefer interpreter to keep talking, disparity between the source text and target text in terms of the number of words rendered, numbers very important to business users, pragmatics
PCC4	more comfortable to listen to, "pollution" in target language, could not bear to listen, up-and-down intonation makes the audience tired of listening, difficult to continue listening, poor delivery difficult for the audience to understand, make people confused, more convincing, lose credibility, examiner point of view: FC outweighs PD, no credibility, very tired of listening to her so did not get the content,
PCC5	convincing tone, does not make people nervous, the audience wouldn't understand, understandable, can get some information, uncomfortable pronunciation and excessive fillers, feel serious interference, mainly consider the audience's perceptions, delivery is more important than accuracy – more from audience point of view

The third conceptual category is Audience Point of View with the conceptual properties shown in Table 5-4. This criterion's conceptual properties overlap substantially with those in the Presentation and Delivery criterion, and most of the overlapping properties are related to the texture of the target language and the way the

interpreters controlled their voices. The examiners specifically stated that they were commenting from the audience's point of view (see Comment 6 in 5.4.2 below). This may be because these conceptual properties are those that may be easily detected by the audience even without understanding the source language.

Some concepts in this category are also linked to Fidelity and Completeness. This is mainly because the audience would not be convinced and continue to listen if the interpretation does not provide credible information to help understand the proceedings at conferences or meetings.

5.4.1 Two requirements from the audience's point of view

The conceptual properties in this criterion are those that the examiners required of the student interpreters to satisfy the audience and fall mainly into two areas: first, to gain the confidence of the audience, and second, to deliver the speaker's message at an acceptable level of faithfulness, so the audience can keep listening to the interpretation. Table 5-4.a presents how the conceptual properties are sorted into the two requirements.

Gain the confidence of the audience

Examiners hold the view that in order to gain the audience's confidence, the interpreter needs to warm-up fast and get into the SI working mode immediately the speech starts. The audience would expect to hear a voice that is professional; the interpretation needs to be concise and coherent, the delivery fluent, natural, and friendly so they can be easily understood. This view echoes Schjoldager's argument that the interpreter's choice of language needs to be adequate and the interpretation needs to be coherent; otherwise, the audience may get irritated and lose interest in the messages, in which case the interpreter's other qualities are irrelevant (Schjoldager, 1995).

Table 5-4.a Requirements of Audience Point of View

Criterion	Requirements	Conceptual properties
Audience Point of View	**Gain confidence**	gain audience's confidence, warm-up fast, delivery impact, more comfortable to listen, willing to keep listening (meaning doesn't really matter), expect to be like a professional, beginner/in-training, unprofessional girlish fillers, breathing control, sigh (microphone usage), concise, fragmented, idiomatic usage, regional usage to target audience, diction, pace, fast=nervous, slow=relaxed, pace, rhythm, tempo, fluency, slow, fast, fluent, mumbling, pauses, polished=easy to understand, less confidence if pauses are long, word choice more suitable to local audience, sighed=loss credibility, more clear and logical from the audience point of view, interpreting is service industry=can't make the audience nervous when listening, alternate point of view: teacher, examiner and audience, feel pressure / no pressure, nervous interpreter makes nervous audience, keep the audience waiting, prefer interpreter to keep talking, "pollution" in target language, could not bear to listen, up-and-down intonation makes the audience tired of listening, difficult to continue listening, more convincing, uncomfortable pronunciation and excessive fillers, mainly consider the audience's perceptions, delivery is more important than accuracy – more from audience point of view
	Faithful delivery	consistent and coherent speech, unfocused, faithfulness to SL, introduce company product, purpose of speech, can get more messages, audience wouldn't know that it's fabrication if the delivery is good but teachers would, startle the audience by saying something out of the context, can follow the interpreter better (less errors and omissions due to overlapping), can hear and understand more clearly, disparity between the source text and target text in terms of the number of words rendered, numbers very important to business users, pragmatics, make people confused, lose credibility, examiner point of view: FC outweighs PD, no credibility, very tired of listening to her so did not get the content, the audience wouldn't understand, understandable, can get some information, feel serious interference,

Faithful delivery of speakers' messages

In spite of the above arguments, Schjoldager emphasized that a disloyal interpreter is unprofessional (ibid). The Audience Point of View criterion observed here also requires an interpretation that carries an acceptable level of faithfulness. But how can someone who does not understand the source speech, a real audience member for instance, distinguish whether or not the interpretation is faithful?

The examiners pointed out that although the audience may not understand the source language, they may still be able to sense that something is wrong or missing in

the interpretation. For example, the interpretation does not match with known facts or violates common sense, or the amount of information gained from the interpretation seems inadequate to make sense of what the speaker is saying. From this perspective, the audience may get an idea of whether or not an interpreter's interpretation output is complete and faithful to the source speech. Previously in 5.3.1 the contextual consistency was discussed as one of the conceptual properties in the criterion of Fidelity and Completeness. Any out of context interpretation is likely to alert the audience, or even make it difficult for the audience to follow. All these may raise doubts in the minds of the audience, making it more difficult for them to keep listening to the interpreter.

5.4.2 Natural delivery vs. faithful interpretation

Most examiners in the five clusters agree on the need to consider the audience's point of view, but as in the previous two categories, there are some variations in the examiners views. Interpreter examiners in PCC3, PCC4 and PCC5 tended to emphasise the Audience Point of View criterion much more than the examiners in PCC1 and PCC2 did, stressing that interpreting is a service industry; therefore, it is important to make the interpretation clear, logical and audience-friendly. Some examiners were insistent on these audience needs. In order to make the point clear they even commented that meaning does not really matter because the audience do not understand the source speech. Comment 6 below illustrates this point of view.

> **Comment 6 (translation)[29]:** I feel that as an audience member first of all you listen to the interpretation…the main prerequisite for the audience to use the interpreting service is that the (interpreted) message must be understandable. […]

[29] **Comment 6** in source text Chinese: 我覺得做一個觀眾來講呢，你首先要聽口譯，他想要用口譯服務的最大一個前提就是，他的訊息我一定要先聽得懂。[…]假設我是觀眾我不會去聽原文…不管他這個訊息是不是符合原來講者的訊息，首先我一定要聽得懂我才會繼續把耳機戴著，繼續往下聽。

> assuming that I'm in the audience and I won't listen to the source speech…it doesn't matter if the messages are in line with the speaker's messages or not, first of all I must understand them so that I will keep wearing the headset and listen.

In other words, some examiners hold the view that delivery is more important than accuracy, that the "feeling" or perception of the customers, i.e. the audience, should be most important when considering the quality of interpreting. From their work experience in the market, for example, they realised that what is regarded as important by the interpreters themselves actually may not be as important to the audience. If the interpreter sounds fearful and unsure, that may be worse than an interpretation that may contain some small errors but is very convincingly delivered.

Of course, it is another matter if the message is completely different from what the speaker says. In arguing for the importance of delivery in interpreting, however, the above view takes Schjoldager's (1995) argument one step further and from the audience point of view puts more weight on the Presentation and Delivery criterion than the Fidelity and Completeness criterion (also see 6.4).

5.4.3 Alternative perspective

There is another interesting concept (or examiner behaviour) in this conceptual category – an alternative perspective when assessing interpreting performances. An examiner may shift the point of view from *teacher* examiner to *audience* examiner. The following two comments illustrate this alternative perspective.

> **Comment 7 (translation)[30]:** Her fluency was really great, but […] I totally cannot accept the accuracy. […] It depends on the perspective that we take as an

[30] **Comment 7** in source text Chinese: 他流暢度真的是很棒，可是[…] accuracy 我真的是完全沒辦法接受，[…]這就是看我們 evaluator 我們是站在什麼身份來看。如果我們是在 evaluate 學生的表現的話，我就會選擇 A。那今天如果我是聽眾的角度來選，我會選擇 C。

evaluator to look at the students' performances. If we are evaluating a student's performance, I will choose A. But if I am making the choice from an audience perspective, I will choose C.

Comment 8 (translation)[31]: I felt that…actually when I first listened my mindset was a bit like a teacher examiner in a professional exam panel assessing the student interpreters. […] However, later I found that some of them…you couldn't tell much difference from their content accuracy. Therefore, I could only do the next thing I could, that is, to simply compare from the audience point of view.

Comments 7 shows that the examiner clearly regards the delivery criterion as more closely related to the view of the audience and the fidelity to that of an evaluator, i.e. an examiner. Therefore, the assessment was conducted with two perspectives; the examiner had to decide which was the more important. In this case, accuracy took priority. In Comment 8, however, the examiner could not distinguish the students' ability levels using the primary criterion, i.e. Fidelity and Completeness, so the judgement was made using the secondary criterion, i.e. from the audience's point of view.

[31] **Comment 8** in source text Chinese: 我會覺得, 其實一開始聽的時候我會覺得我有一點是以老師在給學生考專業考的一個心態在聽， […] 可是後來發覺有幾個人，你沒有辦法在他的內容準確度上做太多的時候，我覺得我只能退而求其次，單純地從聽眾的角度來做比較。

5.5 Interpreting Skills and Strategies

Table 5-5 Conceptual properties of Interpreting Skills and Strategies

Super Examiners	Concepts of the Interpreting Skills and Strategies Category
PCC1	not influenced by source text, paraphrasing, use simpler expressions, approximation of numbers, omissions and correction (when encountering difficulties), summarise, use of visual aid, multi-tasking, know how to distribute efforts, Ear-Voice Span (EVS) range/lagging, short-term memory, speed/pace control,
PCC2	resourcefulness, summarising skills, concise rendition, package different levels of messages, paraphrasing = combine sentences = change meanings, abandon sentences midway not good, self correction, skipping, would rather omit, make up a story, use more neutral statement/expression, say something irreparable, background knowledge support, not sensitive enough in business context, use/not use key words, misuse the slides, anticipation, multi-tasking, overlapping, shorter EVS and sentences, filtering messages, fast processing/translation speed, keep up with the speaker, good timing, good segmentation of sentences/meaning groups, short warm-up time=better skill, not enough processing time, couldn't deal with more complicated messages with strategy, couldn't deal with numbers,
PCC3	Keep up with the speaker, try to translate every word, semi-SI, approximation of numbers, follow the content more closely, summarising, short and condensed, poor multi-tasking, preparation skill, do homework, preparation, background knowledge supports effective delivery, syntactical conversion, message filtering, guessing, paraphrasing, skip, read, ask, rather omit than error, approximation,
PCC4	self-correction, keep up with the speaker, summarising, condense the message, fall too far behind the speaker so lost the messages, prefer omissions than errors, guessing with the few words she understood
PCC5	correction, connecting messages, on-site resourcefulness, making judgement, summary

Examiners also commented on the interpreters' use of strategies. Better strategies normally lead to better interpretation outputs. Therefore, some examiners regard this as an indicator of the level of the students' training and abilities in interpreting, and they take account of the use of interpreting skills and strategies to help them make decisions when comparing the students. Table 5-5 presents the concepts in the Interpreting Skills and Strategies category.

5.5.1 Two types of Interpreting Skills and Strategies

In general, examiners in the five clusters share similar concepts in regard to the use of interpreting skills and strategies, though the variations between the clusters are obvious in terms of the amount and detail of the concepts. The skills and strategies that examiners observed fall mainly into two types: resourcefulness and multi-tasking (Table 5-5.a).

Table 5-5.a Types of Interpreting Skills and Strategies

Criterion	Types	Conceptual properties
Interpreting Skills and Strategies	**Resourcefulness**	on-site resourcefulness, not influenced by source text, paraphrasing, use simpler expressions, approximation of numbers, omissions and correction (when encountering difficulties), summarise, use of visual aid, resourcefulness, summarising skills, concise rendition, package different levels of messages, paraphrasing = combine sentences = change meanings, abandon sentences midway not good, self correction, skipping, would rather omit, make up a story, use more neutral statement/ expression, say something irreparable, background knowledge support, not sensitive enough in business context, use/not use key words, misuse the slides, anticipation, try to translate every word, approximation of numbers, short and condensed, preparation skill, do homework, preparation, background knowledge supports effective delivery, syntactical conversion, guessing, paraphrasing, skip, read, ask, rather omit than error, approximation, connecting messages, guessing with the few words she understood
	Multi-tasking	multi-tasking, know how to distribute efforts, Ear-Voice Span (EVS) range/lagging, short-term memory, speed/pace control, overlapping, shorter EVS and sentences, filtering messages, fast processing/translation speed, keep up with the speaker, good timing, good segmentation of sentences/meaning groups, short warm-up time=better skill, not enough processing time, couldn't deal with more complicated messages with strategy, couldn't deal with numbers, Keep up with the speaker, semi-SI, follow the content more closely, poor multi-tasking, message filtering, making judgement, correction,

Resourcefulness

Being resourceful, the interpreter can use various skills and strategies, such as paraphrasing, summarising, approximation of numbers, to make the message transfer

smooth and easy to understand, which are more related to the language proficiency skills. A competent interpreter can also make use of any visual aids provided to help with understanding the source speech. When encountering difficulties, the interpreter may opt for the reduction strategy or omissions; when realising that a mistake has been made, the interpreter may self-correct where appropriate.

Other skills and strategies include preparation and anticipation with appropriate background knowledge support. For example, some examiners would decide a winner because the student interpreter appeared to have a better background knowledge in business. Comment 9 below is a typical example of such judgement.

> **Comment 9 (translation)[32]:** I don't know these students at all so I'm guessing…the reason why B is better is because she has a better business background than A. My feeling is that there should be three decisive dimensions for [assessing] interpreting: professional subject knowledge, language ability, and interpreting skills; they should be combined together for analysis.

Multi-tasking

In order to implement the above skills and strategies in an effective way, an interpreter needs to have acquired the second type of skill – multi-tasking. According to Gile's Effort Model, interpreters need to be able to balance or manage various efforts when performing multiple tasks at the same time. In the case of simultaneous interpreting, the tasks involved are listening to and analysing the messages, the memory effort, and the production effort, i.e. the delivery of the messages in the target language. If these efforts are not managed well, the simultaneous interpreting is less likely to be successful (Gile, 1995a: 159-185).

[32] **Comment 9** in source text Chinese: 那麼比較好的原因可能是 B，我的猜想喔，因為我完全不認識這些學生，他可能會有商業的背景超過了 A，我感覺口譯應該有三個決定性的 dimensions，一個就是專業知識，一個是語言能力，一個是口譯技巧，我覺得這三者必須要合而為一來探討。

Comment 10 below shows a typical example of an examiner's observation on problems in a student's multi-tasking ability.

> **Comment 10 (translation)[33]:** I'm more concerned that […] she lagged behind the speaker more (than the other students did). She needed to listen more before she could start talking. She could speak in a fast pace, which is not a problem. […] However, she tended to be wordy as she could be relatively fluent when speaking. […] She already needed [more time] to listen more, but still tended to say unnecessary words when she did start talking, which leads to even longer lagging. This may further impair her capacity to listen to what the speaker was saying.

In short, the examiner was saying that the student interpreter could not multi-task very well even though she could speak very fast. When talking, she could not also listen to the speaker. Poor multi-tasking ability usually leads to an unsuccessful performance in simultaneous interpreting.

5.5.2 Views from the interpreter examiners

As mentioned earlier, the variations between the clusters are obvious in terms of the amounts and details of the concepts in this category. For example, PCC2 examiners commented noticeably more on the Interpreting Skills and Strategies than the other examiners did. This is probably due to the fact that the majority of the individual examiners in PCC2, eight out of ten, come from the field of interpreting, and seven of the eight interpreter examiners also teach SI in higher education; hence, these examiners are more used to commenting on students' performances.

---

[33] **Comment 10** in source text Chinese: 我比較會擔心[…]她常常落後比較多，常常會聽比較多才開始說話，那她講話的速度是很快的，這個是沒有問題的，[…]可是可能因為她是比較流利，所以她會講一些多餘的話，[…]她原本就已經快要聽比較多了，結果她可以開始講話的時候又講了沒有必要的話，就表示她會落後更多。這樣有可能她就更不會聽到講者在說什麼。

A PCC2-type examiner tends to give a more detailed account of the skills and strategies that the student interpreters used in SI. In addition to those mentioned earlier in 5.5.1, PCC2 also commented on strategies like using neutral statements and expressions to deal with ambiguous source material, or to cover uncertainties that were due to partial comprehension of the messages by the interpreters. Other strategies include filtering and segmenting sentences and meaning groups in SI, which can help interpreters to deliver a more concise and smoother interpretation of the source speech.

Although there are fewer comments from clusters PCC3, PCC4 and PCC5 (partly because they have few members), these three clusters contributed interesting conceptual properties to this criterion. Both PCC3 and PCC4 emphasised that it is preferable for the interpreters to omit messages rather than to give incorrect ones, which is linked to the concept of "different levels of fidelity" in the Fidelity and Completeness criterion. PCC5 examiners did not elaborate much in this category but emphasized the interpreters' need for on-site resourcefulness and judgement.

5.5.3 Ear-Voice Span (EVS)

Use of interpreting skills and strategies is supported by the multi-tasking ability. However, it is difficult to *directly* measure the multi-tasking ability because the distribution of efforts is in the mind of the interpreter. The examiners can at best *indirectly* measure this ability by observing how an interpreter performs the simultaneous interpreting (SI) task, and then infer from the observation whether or no the interpreter is multi-tasking. "The crucial feature of synchrony in SI is the 'time lag', also know as ***décalage***, between the original speech and the interpreter's output" (Pöchhacker, 2004: 117). This time lag, often referred to as the Ear-Voice Span (EVS), is "one of the few observable variables in SI study (Lee, 2002). Therefore, researchers

regarded EVS as the main indicator of the interpreters' multi-tasking ability because it "extended beyond temporal measurements to the cognitive activity underlying the delay" (Pöchhacker, 2004: 117)[34].

In the context of interpreting assessment, as noticed in this study, examiners also take advantage of observing the EVS to judge how well the student interpreters' SI performances. On one hand, a longer EVS usually means that the interpreter requires better memory retention ability in multi-tasking because the messages need to be retained longer while listening, analysing, remembering, and speaking all at the same time. On the other hand, a shorter EVS is an indication that the interpreter may have a better or faster processing ability in analysing the messages and delivering them in shorter segments. Depending on the pace of the source speech and the complexity of the messages, interpreters fall behind the speaker at various lengths of EVS when simultaneously interpreting. In conjunction with other assessment criteria the examiners observe how the interpreters manage their EVS during the simultaneous interpreting examination to assess whether can effectively multi-task.

In some cases, however, an excessively long EVS does not necessarily mean better multi-tasking, but indicates a lack of listening comprehension, saturation or an overload of the working memory, in which case the interpreter simply stops interpreting. The result is usually obvious omissions of messages.

[34] See Pöchhacker (2004: 117-118) for more reviews on some early studies of EVS in simultaneous interpreting.

5.6 Foundation Abilities for Interpreting

Table 5-6 Conceptual properties of Foundation Abilities for Interpreting

Super Examiners	Concepts of the Foundation Abilities for Interpreting Category
PCC1	Comprehension=accuracy, completely lost, better comprehension, did not grasp, fabricating messages, not comprehension problem, nervous so not understand, long lags = did not hear, slow pace = poor comprehension, (fragmented, incomplete sentences, backtrack repair messages, incoherent) => have severe comprehension problem, misunderstood, same level of comprehension, show consistent comprehension
PCC2	did not understand the message, (most of the time) good listening comprehension, need more training in listening comprehension, excessive fillers (nervousness) impede listening comprehension, not sure if she understood the message, understood the message but did not express well in Chinese, did not hear the speaker due to serious overlapping
PCC3	Didn't hear the important messages, didn't understand the message, didn't understand the speaker, weaker in listening, personality/aptitude should be part of the criteria, use different measures/criteria according to different abilities, common problem/mistake vs. specific/individual problems,
PCC4	Overall understanding of the speech, did not understand, she did not interpret so she did not understand, big comprehension problem, might have understood but could not express clearly, too nervous to listen and understand, working languages, Language A, "pollution" in target language, A-B language combination, English not good enough as B language
PCC5	cannot understand at all, not easy to listen to and understand

When commenting on the students' interpreting performances, the subject examiners also referred to what may be called the foundation abilities for interpreting. Table 5-6 shows the concepts in this category.

5.6.1 Two types of foundation ability

The concepts in this category focus on two types of foundation ability: language ability and personality/aptitude of the student interpreters. In respect to the language ability, the examiners' comments mostly refer to students' listening comprehension of the source speech; some specifically mention target language output ability, i.e.

speaking. This may be due to the fact that most of the student interpreters were interpreting from their second language into their first language so the demand on listening was higher. In addition, most comments on the target language were linked to the Presentation and Delivery criterion. Table 5-6.a shows how these conceptual properties are sorted.

Table 5-6.a Types of Foundation Abilities for Interpreting

Criterion	Types	Conceptual properties
Foundation Abilities for Interpreting	**Personality and aptitude**	nervous so not understand, excessive fillers (nervousness) impede listening comprehension, personality/aptitude should be part of the criteria, use different measures/criteria according to different abilities, common problem/mistake vs. specific/individual problems, too nervous to listen and understand,
	Listening comprehension	Comprehension=accuracy, completely lost, better comprehension, did not grasp, fabricating messages, long lags = did not hear, slow pace = poor comprehension, (fragmented, incomplete sentences, backtrack repair messages, incoherent) => have severe comprehension problem, misunderstood, same level of comprehension, show consistent comprehension, did not understand the message, (most of the time) good listening comprehension, need more training in listening comprehension, not sure if she understood the message, understood the message but did not express well in Chinese, did not hear the speaker due to serious overlapping, didn't hear the important messages, didn't understand the speaker, weaker in listening, overall understanding of the speech, she did not interpret so she did not understand, big comprehension problem, might have understood but could not express clearly, working languages, Language A, "pollution" in target language, A-B language combination, English not good enough as B language

Personality and aptitude

Among the five clusters, only PCC3 clearly mentioned the concepts of aptitude and personality for being an interpreter. The others mentioned it indirectly, such as getting nervous easily, and the ability to remain calm in stressful situations. The reason why this concept was less commented on may be that personality and aptitude is usually one of the *selection* criteria when admitting students into an interpreting training programme. During the interpreting examination, therefore, it may not be appropriate to regard personality and aptitude as an assessment criterion because it is an innate ability, not a

trained ability. However, as personality and aptitude do affect interpreting performance, some examiners mentioned this when they observed excessive nervousness. In Chapter 7, this issue will be discussed in more details.

Listening comprehension

When commenting on the SI performances, the examiners also related student interpreters' listening comprehension closely to the other criteria mentioned in previous sections. When a student interpreter did not grasp the messages, or was even completely lost, the examiner decided that it was due to poor or, lack of, listening comprehension. The inferences were mainly based on the observation of the way the messages were delivered, i.e. Presentation and Delivery criterion, and on the content accuracy. Comments 11 to 14 below illustrate how examiners related the interpreters' performances to their listening comprehension during SI.

> **Comment 11 (translation)[35]:** Well, D is slower, which shows that she sounds struggling, she struggles to listen to and understand the source speech.
>
> **Comment 12 (translation)[36]:** Her (delivery) speed is too slow, which makes you wonder, hey, this (interpreter) may not understand (the source speech).
>
> **Comment 13 (translation)[37]:** Her comprehension, i.e. the part of content accuracy, overall I think is better than D.
>
> **Comment 14 (translation)[38]:** The second point I think she did better is that her pace control is better. Therefore, I think she has more time and capacity to listen, she missed less information.

[35] **Comment 11** in source text Chinese: 那 D 比較慢，顯得他聽起來比較吃力，他聽原文理解的部份比較吃力。
[36] **Comment 12** in source text Chinese:他速度太慢了，會讓你覺得說，耶，這個好像是聽不懂。
[37] **Comment 13** in source text Chinese:他的 comprehension，就是 content accuracy 的部份我覺得整個來說比 D 好。
[38] **Comment 14** in source text Chinese:第二個我覺得他做得比較好的地方，就是他的速度控制地比較好，所以我想他比較有餘裕去聽這個，他比較沒有漏掉什麼東西，

The examiners regarded excessively long EVS or lags as an indicator of the student interpreters' impaired listening comprehension in SI, and commented that it was reflected in their slow and struggling delivery.

Furthermore, some examiners also made a distinction between *not understanding* the message and *not hearing* the message at all. These examiners mostly belong to PCC2. In most cases where the speaker's messages were missed, a PCC2-type examiner would comment that the reason was due to the inadequate use of interpreting strategies or poor multi-tasking ability, so that the interpreter's listening comprehension of the source speech was severely compromised. Unless it could be immediately verified by asking the student interpreters, this was, of course, more of an educated guess, or speculation, based on the examiner's observations and experiences.

As simultaneous interpreting is not a listening test, in some cases after the ten paired comparisons, the examiners were asked to clarify how they made a judgement on the interpreters' listening comprehension. Below are two typical examples of the subject examiners' replies.

> **Reply 1 (translation)[39]:** How would I know? Well, from her interpretation output. […] A paused a lot. She rushed to interpret when she felt that she could understand, but the delivery was in a rush. Because she had to listen to the next sentence, she would pause and then interpret again quickly. This often made her stop in the middle of a sentence, leaving the sentence unfinished. […] or even when she did finish them, the messages were still incorrect.

> **Reply 2 *(in English)*:** I just…again looking at the accuracy or level of accuracy of interpretation. But you know it may not be 100% true, […] because it may be a problem of multi-tasking. You know, some of them are not that good, so given time, their comprehension may be better. […] you can tell from the output.

[39] **Reply 1** in source text Chinese: 怎麼知道啊？從他傳達出來的譯文啊。[…]A 停頓很多，覺得他聽懂的部份他就急急忙忙趕快講，那講出來很趕，因為他忙著要聽下一句，那在聽下一句的時候他又停下來了，然後再急急忙忙地講。那常常講到一半就不見了。他比較多那種半句話，或者即使講出來了，仍然是錯的。

From replies like the above, it is safe to say that a typical examiner in an interpreting examination is looking for an interpreter who shows consistent comprehension of the source speech by observing the interpreter's delivery, interpreting skills, and the accuracy level of the interpretation. These criteria are interrelated and involve some speculations or diagnosis on the examiners' part based on their professional experiences.

5.7 Assessment criteria in the decision-making process

This chapter set out to answer two questions, the first of which is: what assessment criteria are actually used by the examiners in simultaneous interpreting examinations? By examining the examiners' paired comparison comments, five assessment criteria were identified. They are Presentation and Delivery, Fidelity and Completeness, Audience Point of View, Interpreting Skills and Strategies, and Foundation Abilities for Interpreting. The conceptual properties of these criteria are often linked with each other. In other words, examiners are making holistic judgements on student interpreters' performances by considering various assessment criteria at the same time.

Based on these findings, this section attempts to ascertain the relationships between the assessment criteria and the judgement results, i.e. to answer the second question: were the ranking patterns the results of differences in examiners' use of assessment criteria, or did the examiners use similar criteria but on this basis rank students differently?

5.7.1 Primary assessment criteria in paired-comparison decisions

In the interview part of this study, the subject examiners were asked to give the main reason of why they chose the winner (3.2.4.b). Since the "main reasons" were based on the examiners' judgements, their why-the-winner responses usually were in the key words as appeared in the comments like faithfulness, accuracy, delivery, good/poor interpreting skills, etc., which could be matched and sorted into the categories that were emerged from the examiners' own paired comparison comments. They were thus extracted and sorted into the five identified criteria categories. For illustration, below

are samples of examiners' responses in English translation, each representing one identified criterion.

> **Presentation and Delivery**[40]**:** *(The main reasons why you feel B is better are?)* More fluent, higher accuracy rate, the use of terms more accurate.
>
> **Fidelity and Completeness**[41]**:** In comparison, C is much better than D because D has made too many mistakes, including translation errors and number errors […].
>
> **Audience Point of View**[42]**:** The reason why D is worse is that she was slow in getting into the situation [of doing SI] and did not catch the attention of the audience at the very beginning; [she did not] establish trust among the audience.
>
> **Interpreting Skills and Strategies**[43]**:** I feel that E did a better job because she was closer to the so-called simultaneous interpreting. […] In terms of interpreting skills, if I were her teacher, I would feel that E had made progress. […] As for A, she probably was still one step behind so I feel that E is better.
>
> **Foundation Abilities for Interpreting**[44]**:** I think D is probably better than E because I don't think E understood what the speaker was talking about in the first half of the three minute task.

Those main reasons were counted and reported by using their corresponding criteria categories, i.e. "Criteria tallies", as shown in Table 5-7 (also see tables in 5.7.4). The criteria tallies are reported in their percentages in terms of the total number of paired comparisons in each cluster, i.e. "number of decisions" in the table. In the cases where more than one reason was given, the first was recorded for Table 5-7.

Based on the data in Table 5-7, the alpha ICC is calculated. The ICC results indicate that the five super examiners' use of the criteria is highly consistent, individually (single- measures 0.92) and as a whole (average-measures 0.98), when

---

[40] *（那主要的理由你覺得 B 比較好是因為？）*比較流暢、正確度比較高、用詞比較正確。

[41] 比較起來的話，C 比 D 好很多。因為 D 的錯誤非常多，包括誤譯、數字錯誤[…]。

[42] D 比較不好的理由是進入狀態很慢，一開始沒有抓住聽者的心，沒有建立聽者對他的信任感。

[43] 我覺得 E 做得比較好，因為他比較貼近所謂的同步口譯。[…]就做口譯的技巧來講的話，可能如果我是老師的話，我會覺得 E 讓我看到進步。[…]這 A 可能還差一關，這樣的話我覺得 E 比較好。

[44]我想 D 可能比 E 好一點，因為 E 就我聽到的這三分鐘的前面，我就會覺得他可能都不知道講者在講什麼。

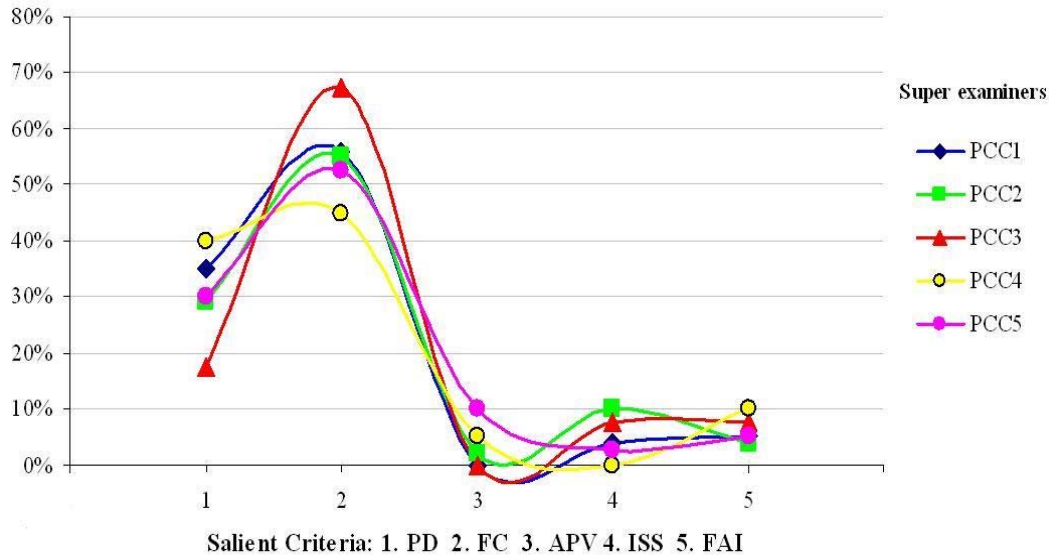
assessing the ten pairs of students. Figure 5-1 shows the line graphs of the super examiners' use of the five criteria. The line patterns of the five super examiners overall are almost identical to one another, though with some wider variations like PCC3.

Table 5-7 Salient criteria used for the PC decisions

Clusters (number of decisions)	Criteria tallies and percentages				
	PD	**FC**	**APV**	**ISS**	**FAI**
PCC1 (98*)	34 35%	55 56%	0 0%	4 4%	5 5%
PCC2 (100)	29 29%	55 55%	2 2%	10 10%	4 4%
PCC3 (40)	7 17.5%	27 67.5%	0 0%	3 7.5%	3 7.5%
PCC4 (20)	8 40%	9 45%	1 5%	0 0%	2 10%
PCC5 (40)	12 30%	21 52.5%	4 10%	1 2.5%	2 5%
Total (298 decisions)	**90** **30%**	**167** **56%**	**7** **2.4%**	**18** **6.1%**	**16** **5.4%**

PD: Presentation and Delivery, FC: Fidelity and Completeness, APV: Audience Point of View,
ISS: Interpreting Skills and Strategies, FAI: Foundation Abilities for Interpreting
*There are only 98 decisions in PCC1 because of tied results in two paired comparisons.

Figure 5-1 Line graph profiles of the PC examiners' salient criteria



Further examining the data in Table 5-7, two primary criteria used in the decision-making process were identified: Fidelity and Completeness (56%) and Presentation and Delivery (30%). The two criteria combined account for 86% of

decisions made in the paired comparisons. These two criteria also fit nicely in the two inner layers – *accurate rendition* and *adequate expression* – of Pöchhacker's (2001) model of the quality standards for interpreting (Figure 2-2).

Next, the question is: can the assessment criteria usage be related and matched to the judgement results? In 4.1.2.b, we have established that the thirty examiners as a group judged Cherry to be the best, followed by Beth, Eileen, Daisy, and with Ally as the worst, i.e. the ranking order on the 30-examiner PC T scale (Figure 4-3.a). Given the two identified primary criteria here, it would be logical to infer that Cherry's interpretation quality was the best in terms of fidelity and delivery of the messages, and Ally's was the worst.

Nevertheless, we also found that the five PC super examiners judged the students with five different ranking patterns (4.5.1). Only PCC1's ranking (Figure 4-5) is the same as the 30-examiner PC T scale (Figure 4-3.a), albeit the distances between the five students are different on the two T scales. In fact, in the five PC cluster T scales (Figure 4-5), Beth came on top three times and Ally left the bottom position twice. That is to say, although the five PC super examiners appear to use the criteria similarly as just discussed above, they judged the students with different ranking patterns, i.e. similar criteria, different judgements.

Variations of conceptual properties among the examiners were also observed in the previous sections (5.2-5.6). Could these variations have contributed to the mismatch between the student ranking results and the examiners' use of the assessment criteria? Will these variations affect changes in overall student rankings? To explore further, sample paired comparisons were selected for investigation by examining the examiners' choice of paired-comparison winners. From Table 4-10 and Appendix D, it was noticed that the thirty examiners' choices of winners differ most in two pairs: Beth-Cherry and Daisy-Ally; in contrast, the choices are most consistent in another two pairs: Ally-Beth

and Cherry-Eileen. Therefore, these four pairs were selected for further examination in more details to find out how the super examiners applied the assessment criteria.

As mentioned above, the examiners were asked to give the main reasons for their choice of winners. Their responses were sorted and presented in Tables 5-8.a, 5-8.b, 5-8.c, 5-8.d, each table representing one of the four paired comparisons selected above. Cronbach's alpha ICC is calculated by using the data in the tables to help with the analysis in the sections below. Since the numbers of decisions are relatively small in each selected paired comparison for robust statistical generalisations, line graph profiles (Figures 5-2, 5-3, 5-4, 5-5) based on their corresponding tables were produced as visual representations of the super examiners' usage of the assessment criteria. The paired-comparison winners in the tables were derived from Appendix D, i.e. the one who received most votes in the pair. The percentages in the tables were calculated based on the criteria tallies by using the method in Table 5-7 above. They indicate how often or how much the super examiners used these criteria based on their reported main reasons for making the decisions.

5.7.2 Inconsistent judgements between clusters

5.7.2.a Similar criteria, different judgement

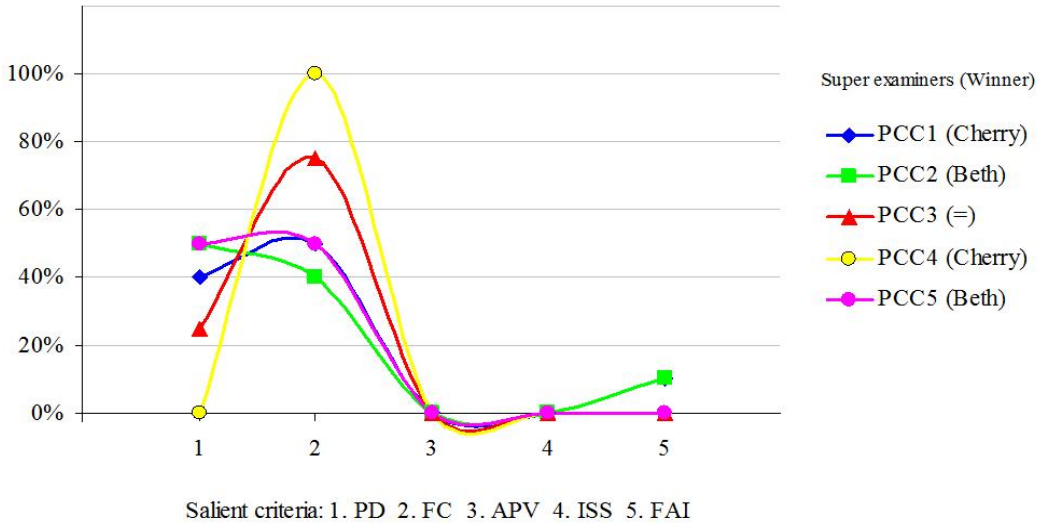
Let us first look at the two pairs, Beth-Cherry and Daisy-Ally, in which the examiners have the least agreement in terms of the paired-comparison winners. Normally, similar criteria should lead to similar judgement results, which is what kept the overall-consistency level high among the thirty examiners' judgements as a group. However, the results of the investigation here show otherwise as in the case of Beth-Cherry comparison (Table 5-8.a and Figure 5-2). They are similar to the observations

Table 5-8.a Salient criteria used for the Beth-Cherry comparison

Clusters (no. of decisions)	Winners	Criteria tallies and percentages				
		PD	FC	APV	ISS	FAI
PCC1 (10)	C	4 40%	5 50%	0 0%	0 0%	1 10%
PCC2 (10)	B	5 50%	4 40%	0 0%	0 0%	1 10%
PCC3 (4)	=	1 25%	3 75%	0 0%	0 0%	0 0%
PCC4 (2)	C	0 0%	2 100%	0 0%	0 0%	0 0%
PCC5 (4)	B	2 50%	2 50%	0 0%	0 0%	0 0%
Total (30 decisions)		12 40%	16 53%	0 0%	0 0%	2 7%

PD: Presentation and Delivery, FC: Fidelity and Completeness, APV: Audience Point of View,
ISS: Interpreting Skills and Strategies, FAI: Foundation Abilities for Interpreting
=: tied, see Table 4-12.

Figure 5-2 Line graph profiles of salient criteria: Beth-Cherry comparison



above in 5.7.1, i.e. the super examiners applied similar assessment criteria, but they might still make different judgement results.

In the Beth-Cherry comparison (Table 5-8.a, Figure 5-2), the super examiners' consistency level of criteria use is excellent as a group of five (average-measure ICC 0.93) and acceptable as individual super examiners (single-measure ICC 0.73), which is similar to the observations on all paired comparisons in 5.7.1. In the Beth-Cherry comparison, the super examiners also showed different judgement results: four super

examiners chose two different winners, and PCC3 could not decide which the better student interpreter was.

When calculating the alpha ICC, the SPSS software also provides Item-Total Statistics, i.e. the item-total correlations (ITC), to show the consistency levels of individual items (here the super examiners) in relation to the group. Inter-item correlations are also calculated, which indicate the consistency levels between two items, i.e. two super examiners. We shall use ITC and the inter-item correlations to further scrutinise the relations between the super examiners' use of assessment criteria and their choice of winners. The statistics are reported in the parentheses below.

ITC shows that in Beth-Cherry comparison, PCC1 (0.95), PCC3 (0.97) and PCC5 (0.88) have excellent and very good consistency levels of criteria usage in the group. However, despite the consistency, the super examiners made three different decisions: two winners and one tied decision (Table 5-8.a). The inter-item correlation between PCC1 and PCC2 (0.96) indicates that the two super examiners are highly consistent with each other in their criteria usage; however, they still picked different winners in spite of the choice being based on a highly consistent criteria usage pattern.

ITC also shows that the criteria-use consistency levels of PCC2 (0.79) and PCC4 (0.74) are acceptable in the group. However, when disregarding the other three examiners, the inter-item correlation between PCC2 and PCC4 (0.48) is unacceptable, which means that they have very different criteria usage patterns (Figure 5-2); they also picked different winners. Since PCC4 has very small counts in the number of decisions, it opens to individual examiners' variation. These findings, however, show that examiners could also apply the criteria differently and make different judgements, which the Daisy-Ally comparison illustrates very clearly below.

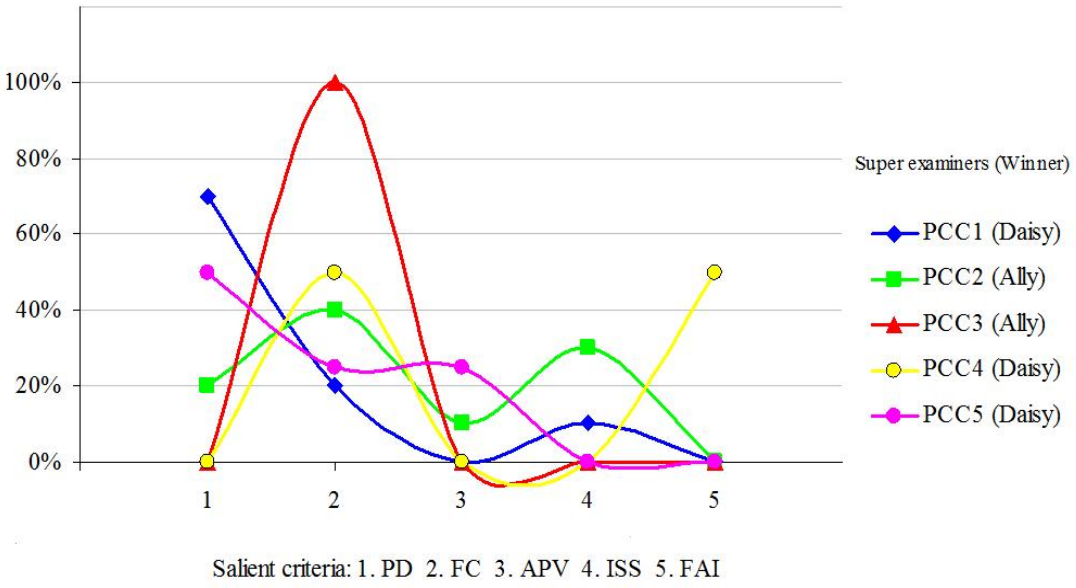
5.7.2.b Different criteria, different judgement

Table 5-8.b Salient criteria used for the Daisy-Ally comparison

Clusters (no. of decisions)	Winners	Criteria tallies and percentages				
		PD	FC	APV	ISS	FAI
PCC1 (10)	D	7 70%	2 20%	0 0%	1 10%	0 0%
PCC2 (10)	A	2 20%	4 40%	1 10%	3 30%	0 0%
PCC3 (4)	A	0 0%	4 100%	0 0%	0 0%	0 0%
PCC4 (2)	D	0 0%	1 50%	0 0%	0 0%	1 50%
PCC5 (4)	D	2 50%	1 25%	1 25%	0 0%	0 0%
Total (30 decisions)		11 37%	12 40%	2 7%	4 13%	1 3%

PD: Presentation and Delivery, FC: Fidelity and Completeness, APV: Audience Point of View,
ISS: Interpreting Skills and Strategies, FAI: Foundation Abilities for Interpreting.

Figure 5-3 Line graph profiles of salient criteria: Daisy-Ally comparison



The Daisy-Ally comparison (Table 5-8.b) is an interesting pair showing how the five super examiners varied criteria usage patterns when making their judgements. In this paired comparison, the super examiners' consistency level is poor as a group (average-measure ICC 0.54), and unacceptable as individual super examiners (single-measure ICC 0.19). Figure 5-3 shows the crisscrossing line profiles of the five super

examiners, which indicate inconsistent use of the assessment criteria when comparing the two student interpreters. The decisions of the super examiners were also split between Ally and Daisy as shown in Table 5-8.b, which means that the five super examiners made different judgements based on different usage patterns of the assessment criteria.

Considering the two larger super examiners PCC1 and PCC2 (which contain twenty individual examiners) between the two paired comparisons (Tables 5-8.a and 5-8.b), their patterns of criteria usage and judgement results also vary between comparisons. In the Beth-Cherry comparison, both super examiners use the two primary criteria similarly, focusing relatively evenly on PD (PCC1 40%, PCC2 50%) and FC (PCC1 50%, PCC2 40%), but they made different judgements in the paired comparison. In contrast, the two super examiners PCC1 and PCC2 apply the PD criterion (70%, 20%) and the FC criterion (20%, 40%) differently, and picked different winners in the Daisy-Ally comparison.

Overall from the above analysis, it appears that the examiners' usage of assessment criteria may vary widely when judging student interpreters. There is no prevailing pattern of the criteria usage in relation to the judgement results, which can apply to the super examiners' decisions. The super examiners may pick the same winners even when they used different patterns of criteria, and when the criteria usage is similar, inconsistent judgement results still appeared.

Having said that, what happens when the examiners make consistent judgement? As mentioned earlier, normally, similar criteria should lead to similar judgement results, but do the super examiners really apply the assessment criteria in similar ways when they make consistent judgement? To find out, next we shall analyse the paired comparisons of Ally-Beth and Cherry-Eileen, i.e. the two pairs in which the examiners have picked the same winners.

5.7.3 Consistent judgements between clusters

This section analyses the two pairs that have the most consistent judgements among the examiners: Ally-Beth and Cherry-Eileen. The super examiners all chose the same winners in these two pairs, i.e. Beth and Cherry respectively. In fact, 28 out of the 30 examiners picked the same two winners (see Appendix D). PD and FC are still the primary criteria (more than 86% combined) in these two paired comparisons.

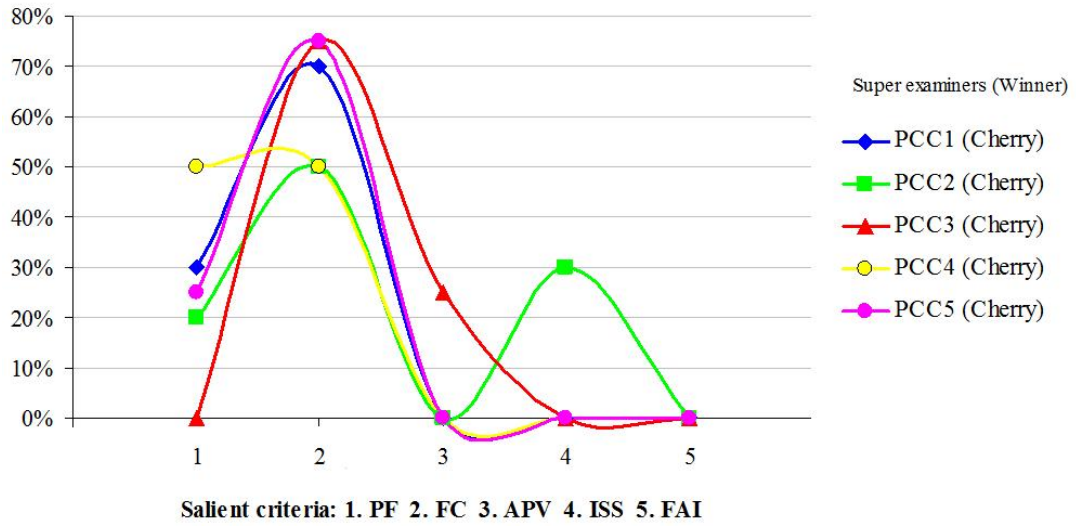
5.7.3.a Similar criteria, similar judgements

Table 5-8.c Salient criteria used for the Cherry-Eileen comparison

Clusters (no. of decisions)	Winners	Criteria tallies and percentages				
		PD	FC	APV	ISS	FAI
PCC1 (10)	C	3 30%	7 70%	0 0%	0 0%	0 0%
PCC2 (10)	C	2 20%	5 50%	0 0%	3 30%	0 0%
PCC3 (4)	C	0 0%	3 75%	1 25%	0 0%	0 0%
PCC4 (2)	C	1 50%	1 50%	0 0%	0 0%	0 0%
PCC5 (4)	C	1 25%	3 75%	0 0%	0 0%	0 0%
Total (30 decisions)		7 23.3%	19 63.3%	1 3.3%	3 10%	0 0%

PD: Presentation and Delivery, FC: Fidelity and Completeness, APV: Audience Point of View,
ISS: Interpreting Skills and Strategies, FAI: Foundation Abilities for Interpreting.

Figure 5-4 Line graph profiles of salient criteria: Cherry-Eileen comparison



Let us first look at the Cherry-Eileen comparison (Table 5-8.c), the alpha ICC suggests that the super examiners' consistency level is excellent as a group (average measure 0.94), and also acceptable as individual super examiners (single measure 0.77). Figure 5-3 shows the line graphs of the super examiners' use of the five criteria. In this pair, the item-total correlations (ITC) show that PCC1 (0.99) and PCC5 (0.99) are most consistent in the group, whose line graphs are almost identical to each other; whereas PCC2 (0.78), PCC3 (0.75) and PCC4 (0.77) have slightly different line graphs, i.e. different usage of criteria, but they are still within acceptable levels. Their line patterns in general, therefore, are identical to one another, but with some variations. This shows the normal situation where consistent use of assessment criteria results in consistent judgement result, all super examiners picked Cherry as the winner.

5.7.3.b Different criteria, similar judgements

For Ally-Beth comparison (Table 5-8.d), the five super examiners used the assessment criteria at an acceptable consistency level as a group (average-measure ICC 0.72), but individually the consistency level is unacceptable (single-measure ICC 0.33). ITCs show that PCC1 (0.89) is the most consistent in the group; however, all the other four super examiners – PCC2 (0.50), PCC3 (0.54), PCC4 (0.00), PCC5 (0.63) – apply the assessment criteria in very different ways. These differences are shown clearly in the line profiles in Figure 5-5, which appear markedly different from one another. In other words, in the Ally-Beth comparison, the super examiners all chose the same winner (i.e. Beth), but the decision was made by using the assessment criteria with different weightings.

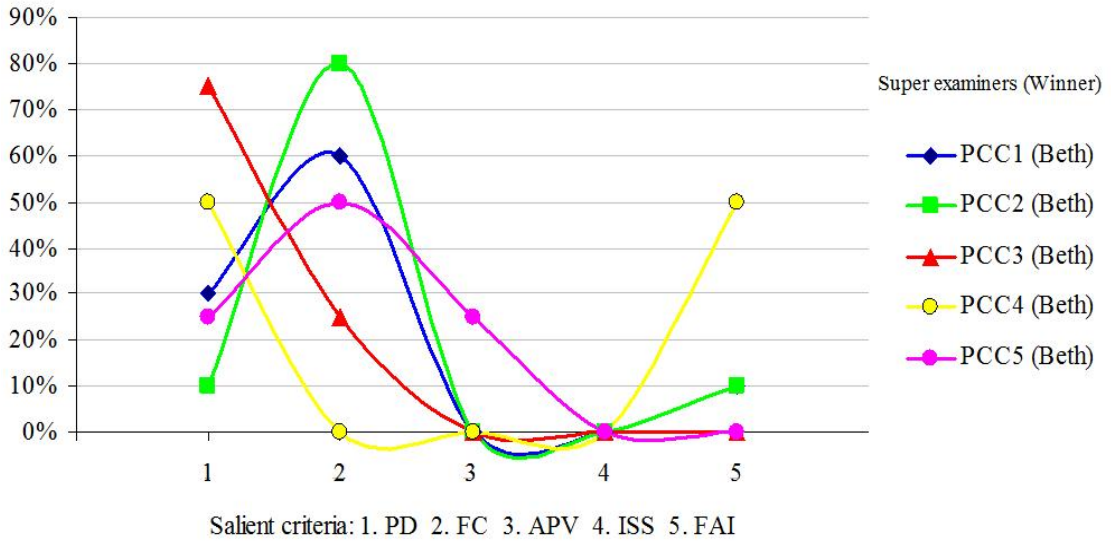
The interesting point is: the above differences in the use of assessment criteria appear to have no effect on the examiners' choice of winner in the Ally-Beth comparison; they all picked the same one despite the differences. One possible

Table 5-8.d Salient criteria used for the Ally-Beth comparison

Clusters (no. of decisions)	Winners	Criteria tallies and percentages				
		PD	FC	APV	ISS	FAI
PCC1 (10)	B	3 30%	6 60%	0 0%	0 0%	1 10%
PCC2 (10)	B	1 10%	8 80%	0 0%	0 0%	1 10%
PCC3 (4)	B	3 75%	1 25%	0 0%	0 0%	0 0%
PCC4 (2)	B	1 50%	0 0%	0 0%	0 0%	1 50%
PCC5 (4)	B	1 25%	2 50%	1 25%	0 0%	0 0%
Total (30 decisions)		9 30%	17 57%	1 3%	0 0%	3 10%

PD: Presentation and Delivery, FC: Fidelity and Completeness, APV: Audience Point of View,
ISS: Interpreting Skills and Strategies, FAI: Foundation Abilities for Interpreting.

Figure 5-5 Line graph profiles of salient criteria: Ally-Beth comparison



explanation may be that the differences in the ability levels of the students in the pair were so great that the winners were obvious regardless of the different use of the assessment criteria. Nevertheless, the fact that the super examiners' use of assessment criteria varies even when looking at the same students raises concerns about potential judgement inconsistencies. It is also possible that, when assessing student interpreters, the examiners followed some unknown judgement approaches or patterns. Since the super examiners are made of individual examiners, we shall further explore this by looking at the judgements of individual examiners.

5.7.4 Judgements of individual examiners within the clusters

The super examiners are made of individual examiners whose judgement patterns, i.e. rankings of students, are more similar to each other than to examiners' judgement patterns in other super examiner group, which was identified from the results of the cluster analysis (4.3). To understand the assessment behaviours of the super examiners better, this section will examine the individual examiners *within* the super examiners. It would be interesting to establish whether the above variations of assessment criteria usage also exist among the individual examiners, with similar criteria resulting in different winners and/or different criteria resulting in the same winner.

Table 5-9.a PCC1 examiners' salient criteria in the Eileen-Beth comparison

Examiner	Winner	Salient criteria used
R1	**E**	Not much different. E slightly better in terms of *completeness*.
R30	**E**	Difficult to compare. E slightly better in *fidelity*.
R28	**E**	E: better *delivery*; B not convincing due to her childish voice.
R16	**E**	E: better *fidelity* (examiner chose B at first)
R15	**B**	Similar in accuracy. B better in *delivery*.
R27	**B**	B: better *fidelity,* completeness, and strategy (numbers)
R3	**B**	B: better *delivery*, fidelity and completeness
R7	**B**	B: better *completeness* and terms
R11	**B**	B: better *fidelity* (main ideas), less meaning errors
R29	**Tied**	First thought B was slightly better, but decided that the two were equal

Note: In the cases where more than one reason was given, the first one mentioned by the examiner is considered to be the salient criterion.

Let us begin with the ten examiners in PCC1, who are mostly non-interpreters (Table 4-9.a) and the most consistent group of individual examiners in terms of student ranking patterns (Table 4-7.a). The PCC1 examiners' judgements are quite consistent on all paired comparisons except the Eileen-Beth comparison (Appendix D), which will be scrutinised here. Table 5-9.a shows the main criterion given by the examiners and the selected winners in this pair. The main criteria are FC (R1, R30, R16, and R27, R7, R11) and PD (R28, R15, R3) as indicated in italics in the table.

However, these examiners picked different winners based on the same criteria. R1, R30, R16 felt that Eileen delivered more relevant key messages, but R27, R7 and R11 commented that Beth's interpretation was more faithful and complete. These six examiners reported using the same FC criterion but made contradictory judgement on the same two students. The other three examiners (R28, R15, R3) who reported PD as the decisive criterion also made contradictory judgements. Examiner R28 felt that Beth's delivery was not convincing as an interpreter and sounded "like a child talking", but R3 and R15 liked Beth's "steady and evenly paced" delivery.

From the other perspective, where examiners picked the same winners, four examiners (R1, R30, R28, R16) judged Eileen to be better, and six (R15, R27, R3, R7, R11) selected Beth. However, the main criteria reported within these two same-winner sub-groups were different (Table 5-9.a), i.e. different criteria, same judgements.

Examiner R29 considered both criteria FC and PD at great length, saying that Beth was more consistent throughout her interpretation, and that Eileen's performance was mixed with very good and poor segments, who was less consistent but had good potential to train well. Therefore, R29 decided that, overall, the two students' interpreting performances were equal.

It is also worth noticing that R15 commented that the fidelity level was similar between Eileen and Beth so her final decision was made on the delivery criterion (Table 5-9.a). That is, the examiner opted for another criterion when one primary criterion could not separate the two students' interpreting performances. Other examiners also used the similar approach of switching criteria. This assessment behaviour will be reported and discussed with details in Chapter 6.

Within the PCC1 super examiner, the situation where different criteria resulted in the same winner can also be illustrated in another paired judgement, such as in the Beth-Cherry comparison, in which all PCC1 examiners agreed on the same winner

Cherry. This time, in Table 5-9.b, we see that the individual PCC1 examiners' salient criteria range from Foundation Abilities for Interpreting (comprehension), to Fidelity and Completeness, and to Presentation and Delivery. Interpreting Skills and Strategies were also mentioned but not as the first criterion.

Table 5-9.b PCC1 examiners' salient criteria in the Beth-Cherry comparison

Examiner	Winner	Salient criteria used
R1	**C**	C: better *comprehension*, but not all that faithful.
R30	**C**	C: better *fidelity*, get more important information
R28	**C**	C: better *delivery*
R16	**C**	C: better *delivery* (target language more concise)
R15	**C**	C: better *accuracy* and delivery
R27	**C**	C: better *fidelity*, completeness and delivery
R3	**C**	C: better *delivery*, fidelity/completeness, and strategy
R7	**C**	Difficult to compare. C: better *delivery* and strategy
R11	**C**	C: better *fidelity*, completeness and strategy (short lag)
R29	**C**	C: better *completeness* and delivery

Note: In the cases where more than one reason was given, the first one mentioned by the examiner is considered to be the salient criterion.

Now, let us look at the same pair in PCC2, which is a ten-examiner cluster whose individual examiners are mostly interpreters (Table 4-9.a). Unlike in PCC1, the Beth-Cherry comparison is where the PCC2 examiners' decisions were most divided; four examiners picked Cherry and six examiners picked Beth (Table 5-9.c).

Table 5-9.c PCC2 examiners' salient criteria in the Beth-Cherry comparison

Examiner	Winner	Salient criteria used
R4	**C**	C: better *fidelity*, completeness, and delivery
R21	**C**	C: better *completeness* and logic
R8	**C**	C: processing speed is faster (B omitted more), better *delivery*
R13	**C**	C: good *comprehension* and delivery, more complete
R5	**B**	C: strange Chinese *delivery*, more meaning errors
R23	**B**	B: better *delivery*; C too vague
R6	**B**	B: less *major errors*; C: better delivery, but some key errors.
R25	**B**	B: better in *fluency* and completeness; C: delivery problem
R20	**B**	B: better *fidelity*, many omission but minor; C too many errors.
R2	**B**	B: better *delivery*, fidelity; C longer warm up, many pauses

Note: In the cases where more than one reason was given, the first one mentioned by the examiner is considered to be the salient criterion.

Those PCC2 examiners who picked Cherry commented that her interpretation was very faithful, complete and that her delivery was very steady and evenly paced, saying that Cherry did a better job in terms of information processing, and the main reason why she chose Cherry was because her processing speed was faster[45], which led to a better delivery. An examiner even said that there was not much fault to pick in Cherry's interpreting performance, and that all Cherry needed was more practice[46].

On the contrary, the other six PCC2 examiners chose Beth as the winner because they had less confidence in Cherry, saying that her interpretation was vague and contained serious mistakes. One examiner even said that Cherry was wrong "from beginning to the end" with only one or two correct items of information[47]. Some also commented that some of Cherry's Chinese expressions were a bit strange; only R6 felt that Cherry's delivery was better, but still contained more serious mistakes than Beth.

Here we see a strong contrast between the judgements of the two sub-groups of PCC2 examiners. While one sub-group judged Cherry to be very faithful, the other sub-group judged otherwise. Again, this is a familiar pattern we have seen, i.e. similar criteria resulting in different judgements. In fact, the results and comments of the other paired comparisons reveal similar situations of contradictory judgements on the same students' performances and can be found within all of the five clusters.

Of course, we do not expect every examiner to think or behave exactly in the same way, but if we can understand why there are differences in the examiners' judgements, it will help to improve interpreting examinations. It would be especially interesting to find out, for example, why there are such strongly contrasted opinions between the two sub-groups of PCC2 examiners on Cherry's interpreting performance (see Chapter 6).

[45]R8's comment: 在資訊處理上我覺得 C 比較好。主要的理由就是我覺得她處理的速度比較快。
[46]R13's comment: 我覺得 C 沒有什麼好挑惕的，他可能需要的就是再多一些練習。
[47] R20's comment: C 給的 information 錯誤率真的太高了。[…]從前面一開始就從頭錯到尾。我聽到正確的跟本就只有一兩個地方是正確的。

At the beginning of 5.7, we stated that the objective of this section is to ascertain the relationships between the assessment criteria and the judgement results. Given the analysis above, therefore, a quick answer is: we cannot be sure about the relations between the judgement patterns and the examiners' use of the assessment criteria. They are, to use Sawyer's word again, "fuzzy", which will be explored further later in Chapter 6 by looking at the examiners' assessment behaviours.

5.8 Summary discussion

In this chapter we set out to explore and answer two questions: (1) what are the assessment criteria that are actually used by the subject examiners in simultaneous interpreting exams? and (2) are the ranking patterns of the student interpreters' performances the results of differences in examiners' assessment criteria, or do they use similar criteria but, on this basis, rank students differently?

In this chapter, five categories of assessment criteria have been identified to answer the first question. The assessment criteria are Presentation and Delivery (5.2), Fidelity and Completeness (5.3), Audience Point of View (5.4), Interpreting Skills and Strategies (5.5), and Foundation Abilities for Interpreting (5.6). These identified assessment criteria are also useful as a base to formulate the test construct of interpreting, which is still underdeveloped (2.3.2.c and 2.4.2.a). As each criterion contains various properties that the subject examiners distinguish when assessing the student interpreters, this means that they are useful for implementations in a test. In Chapter 7, I will further discuss the relations of these criteria and put them into the context of the test construct of interpreting. Below is a brief summary of the assessment criteria properties.

For Presentation and Delivery, the examiners mainly look at three aspects of the interpretation delivery: acoustic, word/phrase and flow of information. For Fidelity and Completeness, there are three areas of focus: content accuracy, speaker intention, and contextual consistency. For Audience Point of View, an interpreter needs to gain the confidence of the audience and deliver the speaker's message at an acceptable level of faithfulness. For Interpreting Skills and Strategies, the examiners look for two types of ability: resourcefulness and multi-tasking. Being resourceful includes two types of ability. One type is the ability to use skills and strategies such as paraphrasing,

summarising, skipping, self-correction, background knowledge and anticipation; the other type is the multi-tasking ability that supports the above interpreting skills. The multi-tasking ability can be observed by looking at the way interpreters manage their Ear-Voice-Span (EVS) lags. Finally, examiners also look at an interpreter's Foundation Abilities for Interpreting, focusing mainly on listening comprehension, though the aptitude and personality of an interpreter were also mentioned.

Many of the criteria properties in the five categories were also found to be closely related to one another and difficult to judge separately. Therefore, when assessing student interpreters, the examiners may make judgements in a holistic manner with different weightings of the assessment criteria when comparing the students' performances.

Among the five categories of criteria, Fidelity and Completeness (56%) and Presentation and Delivery (30%) combined account for 86% of the 298 decisions made in the paired comparisons (Table 5-7). Therefore, they can be regarded as the primary criteria when the examiners first make paired comparison decisions (5.7), which also fall nicely in the two core layers – *accurate rendition* and *adequate expression* – of Pöchhacker's (2001) model of the quality standards for interpreting (Figure 2-2).

The other criteria, such as the Audience Point of View and the Interpreting Skills and Strategies, may fit into the two outer layers of Pöchhacker's (2001) quality model: *equivalent effect*, and *successful communicative interaction*. However, these criteria and standards are more difficult to operationalise in an interpreting examination. This is not only because of the subjective judgements of the examiners, but also because of 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 in this study relied more on the two primary criteria, i.e. the two core standards, to assess the student interpreters in the simulated examinations.

In order to answer the second question, i.e. ascertain the relationship between the ranking patterns and the use of assessment criteria, the salient criteria of the judgement results of the super examiners were cross-examined (5.7). The answer to the second question is yes and no. Yes, the ranking patterns of the student interpreters' performances result 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 many variations in the assessment approach that the examiners took were observed. We were unable to identify a prevailing pattern of the examiners' use of the assessment criteria in relation to their judgements.

It was found that the judgement approach of the super examiners varies widely in terms of their usage of the assessment criteria. They may apply similar criteria but make contradictory judgements, or use different criteria but still pick the same winners. In some cases, all the criteria used and the judgements made differed between the examiners. This variation in the use of criteria and the corresponding judgement results appears in both the super examiners (5.7.2, 5.7.3) and the individual examiners (5.7.4).

Nonetheless, there are certain characteristics in the way super examiners apply the criteria. For example, the PCC2 super examiner seems to pay more attention to the Interpreting Skills and Strategies, which may be due to the interpreter background of its members. When scrutinising the comments of the thirty examiners to identify the assessment criteria, factors were also noted that may affect the examiners' judgement behaviours (5.1.2).

In next chapter, therefore, we shall explore those factors in order to clarify what the reasons are for the subject examiners' inconsistent judgement patterns and their use of assessment criteria that we have seen in this chapter.

CHAPTER 6

Examiners' Assessment Behaviours

6.0 Introduction

In investigating how individual examiners make their judgments during simultaneous interpreting examinations, fluctuations in the examiners' judgements on the five student interpreters were observed (Chapter 4), salient criteria used by the examiners were also identified, but the use of the criteria varied in relation to the judgement results (Chapter 5). These different approaches to judgement may be related to the different assessment behaviours of the examiners, which were noted when exploring the examiners' paired-comparison comments (5.1.2). The assessment behaviour of an examiner cannot be regarded as an assessment criterion because it is not an observation of the students' performances. However, assessment behaviour may still affect the examiner's judgement, and may be the reason for the inconsistencies that were observed in the previous chapters.

This chapter reports and explores the conceptual properties of the examiners' assessment behaviours to answer the final research question: what lies behind consistencies and inconsistencies in judgements among examiners? Preliminary discussions will be based on the concepts identified in Table 6-1 and supported by sample comments and research notes made during the field study: the simultaneous interpreting examination simulation[48].

[48] Noticing some different examiner behaviours during the examination simulation, the researcher made notes about those behaviours for reference. See Appendix C for a sample of the notes.

6.1 Conceptual Properties of Examiner Behaviour

Table 6-1 Conceptual properties of Examiner Behaviour

Super Examiners	Concepts of the Examiner Behaviour Category
PCC1	notes on scripts, examination recordings, attention, examiner memory lag, pay attention to EVS lags, bias, accent, know students, personal preferences, guessing comprehension, interpreting strategies, interpreter preparation, interpreter tired, training levels, judgement pattern, Fidelity/Completeness/ Delivery approach, marking strategy, quick/slow decisions, weightings of criteria, quality consistency, warm-up time, look for potentiality, from past experiences as the audience, difficult to decide, reverse decision
PCC2	notes (with/without scripts), recording reviews, examination script (slide), forgot the wording but knew it's wrong, give me the script because I forgot, first impression not good due to fabrication, she might have said it and I didn't hear it, guessing interpreting strategies, not many lines on notes so she might not have made serious mistakes, personal preferences, couldn't stand fillers, being subjective, could not tell due to regional differences, better background knowledge, not enough training in numbers, give student suggestions, she didn't hear the number but felt that…(guessing), I guess she noticed a logical error…, I feel that she was summarising and not doing SI, if I could not hear speaker how could she hear it (multi-tasking), doesn't make much sense commenting on too much details (focusing on business sense, etc.), I was too nervous when I first listened to the interpretations, marking strategy, FCD approach, become better and better vs. poor interpretation throughout, do less damage, I didn't hear clearly but I felt she missed a lot, overall is good, primacy/recency effect, different impressions between the first and second reviews, reverse judgement
PCC3	note (with/without script), didn't write down, I don't know if she made the same mistake, anticipate interpreter to perform better, look for potential=give more training, criteria priority (accuracy cover rush delivery), aptitude vs. delivery/accuracy,
PCC4	review recordings, notes, guessing the interpreters' country or origin, I forgot, less dangerous = less errors, did not write it down, did not hear clearly why, can't be bothered to listen because her interpretation was all wrong – definite fail, my impression, overall trainable, more complete more errors, lost a lot of messages but less errors, prefer omissions than errors,
PCC5	noting errors on the script, guessing possible causes, problem less serious, primacy/recency effect, negative impression from the interpreter's booth manner – use of microphone, not sure in some parts, influenced by interpreter's background – word choice, judge by personal impression, delivery is more important than accuracy, more from audience point of view, consider on-site situation, judging from impression/from notes

Table 6-1 contains the conceptual properties related to the examiners' assessment behaviours. These concepts were extracted from the examiners' verbal comments using the coding process explained in 5.1 Coding Process of Interview Data.

Some behaviour is easy to observe, and can be referred to as the examiner's external behaviour, for example using the assessment tools. Other behaviour, however, is less straightforward and will only be inferred in this study by analysing the

Table 6-1.a Types of Examiner Behaviour

Criterion	Types	Conceptual properties
Examiner Behaviour	External Behaviour	**the use of assessment tools** notes on scripts, examination recordings, notes (with/without scripts), recording reviews, examination script (slide), give me the script because I forgot, not many lines on notes so she might not have made serious mistakes, note (with/without script), didn't write down, review recordings, notes, did not write it down, noting errors on the script, , judging from notes
	Internal Behaviour	**a general judgement approach (FCD approach)** marking strategy, Fidelity/Completeness/Delivery approach, from past experiences as the audience, difficult to decide, reverse decision, criteria priority (accuracy cover rush delivery), **examiner attention** attention, examiner memory lag, pay attention to EVS lags, forgot the wording but knew it's wrong, she might have said it and I didn't hear it, give me the script because I forgot, , I was too nervous when I first listened to the interpretations, I didn't hear clearly but I felt she missed a lot, overall is good, I don't know if she made the same mistake, I forgot, can't be bothered to listen, my impression, did not hear clearly why, not sure in some parts, judge by personal impression, **examiner bias** bias, accent, know students, personal preferences, first impression not good due to fabrication, personal preferences, couldn't stand fillers, being subjective, could not tell due to regional differences, primacy/recency effect, different impressions between the first and second reviews, reverse judgement, guessing the interpreters' country or origin, can't be bothered to listen because her interpretation was all wrong – definite fail, influenced by interpreter's background – word choice **professionally-referenced standards** guessing comprehension, interpreting strategies, interpreter preparation, interpreter tired, training levels, judgement pattern, quick/slow decisions, weightings of criteria, quality consistency, warm-up time, look for potentiality, guessing interpreting strategies, better background knowledge, not enough training in numbers, give student suggestions, she didn't hear the number but felt that…(guessing), I guess she noticed a logical error…, I feel that she was summarising and not doing SI, if I could not hear speaker how could she hear it (multi-tasking), doesn't make much sense commenting on too much details (focusing on business sense, etc.), become better and better vs. poor interpretation throughout, do less damage, anticipate interpreter to perform better, look for potential= give more training, aptitude vs. delivery/accuracy, less dangerous = less errors, overall trainable, more complete more errors, lost a lot of messages but less errors, prefer omissions than errors, guessing possible causes, problem less serious, negative impression from the interpreter's booth manner – use of microphone, delivery is more important than accuracy, more from audience point of view, consider on-site situation

examiner's comment. This is because the judgement process and various considerations during the decision-making process are in the mind of the examiners. These judgement processes will be referred to as the examiner's internal behaviour. Table 6-1.a presents how the conceptual properties in Table 6-1 are sorted into the various types of examiners' assessment behaviours. The conceptual properties are a mixture of both key words and phrases of real extracts of the examiners' verbalisations. We shall use Table 6-1.a as a framework of discussion in this chapter.

6.1.1 External behaviour – the use of assessment tools

This study asked the examiners to assess the student interpreters in their normal way imposing no procedure in terms of the usage of assessment criteria and a rating scale[49]. In addition to the use of criteria as identified in Chapter 5, variations also were observed in the way the examiners used the assessment tools, including the use of the speech script and the examination recordings.

Some examiners chose to use the speech script and some did not refer to it, but only listened to the source speech. When the speech script was used, different examiners used it in different ways. For example, some examiners made notes on the script whereas the others just read it as they listen to the students' interpretation. As for the use of examination recordings, during the paired comparisons some examiners would ask to review the examination recordings, whereas the others went through the viewing process relatively quickly without reviewing any of the performances.

This study observed two typical approaches by examiners to familiarise themselves with the source speech during the briefing session before the examination simulation

[49] No assessment criteria or rating scale are provided in this study to the participant examiners. The researcher only acted as a facilitator to administer the examination simulation and the paired comparison interview (see 2.5.2, 2.5.3, 3.2.4).

(3.2.4.a). The difference between the two approaches was especially marked between interpreter and non-interpreter examiners. A non-interpreter examiner was more likely to simply listen to the speech with the script at hand, and make notes of the student interpreters' performances directly onto the script. In contrast, before listening to the student interpreters, an interpreter examiner would often ask to listen to the source speech and simultaneously interpret by whispering without referring to the script. Then, they took notes when listening to the students' interpretation. This type of examiner usually referred to their notes, not the speech script, when assessing the student interpreters. Seven of the nine subject examiners, who did not use the script, were interpreter examiners (Table 3-3).

The practice of examiners trying out the interpreting task before they assess students' performances is not uncommon at a professional interpreting examination (Yang, 2000: 162). The main purpose of doing so is to make sure that the difficulty level of the task is appropriate, and so that the examiners have a better idea of where the difficulties of the task may be. Nonetheless, how this pre-test practice would contribute best to the validity and reliability of the interpreting examination task needs further consideration, which will be discussed in Chapter 7.

In addition, we noticed that some examiners could be very quick in making their judgements, whereas others needed more time before reaching a decision. The time the thirty examiners needed to complete the ten paired comparisons ranged between twenty-five and ninety minutes. This time difference may have resulted from the assessment strategies or judgement approaches that the examiners employed. In this study, some of the examiners reversed their decisions a few times before making final ones. It seemed that those examiners who had a set of assessment strategies or approaches with which they were familiar could make quicker judgements.

6.1.2 Internal behaviour – a general judgement approach

It is more difficult to observe internal behaviours, i.e. how people think. One of the widely used methods for psychologists "to explore the previously inaccessible domains of cognitive processing" and to analyze human thoughts, is verbal report analysis (Kucan and Beck, 1997 in Whittington, López, Schley, & Fisher, 2000). Just like expressing ideas and emotions, people can verbally report what they are aware of when performing a task. According to the theory of verbal protocols (Ericsson & Simon, 1980, 1993), when performing a task – mental or physical – people may temporarily store their thoughts of the processes in their working memory, and can report about the components of high level mental processes, i.e. articulate their thinking, that leads to the solution of a problem. Analysing such verbal reports may help researchers to understand how people think in relation to the task that they do.

Conceivably, the act of verbal report (or thinking aloud) may alter the thinking being reported, which may in turn lead to degrading or distorting the main task being performed. Ericsson and Simon (1993) argued that this may not be the case. Although thinking aloud may slow down the task being performed, it should not change fundamentally if the task is primarily verbal, such as only verbalising the content of working memory, and if the person is not asked to explain or evaluate his or her thinking. Ericsson and Simon tested the validity of this argument and found that the act of introspection in their experiments did not affect subjects' mental processes: subjects go through the same steps whether they concurrently describe what they are doing, retrospectively describe it, or do neither, which suggests that introspection can be practiced in reliable ways as a research method (Ericsson & Simon, 1980, 1993).

There are different types of introspective verbal reports, and the simplest and most natural type is *descriptive introspection* (Farthing, 1992). In such verbal report, people

describe their conscious experience in natural language terms, such as what I perceive, think, or feel. This kind of verbal report concerns meaningful events, objects, people, and thoughts about them rather than abstract generalizations or unnatural analyses of the tasks being performed (ibid). In this study, the interview comments of the participant examiners belong to this type of descriptive introspection. The examiners were asked to verbalise their judgement process while comparing the students' interpreting performances, i.e. a concurrent introspective verbal report of their thoughts. The examiners were *not* asked to evaluate their own judgement approach or the assessment criteria being used, but only to describe them as it happens (see 2.5.3.b and 3.2.4.b). Through verbal report analysis and the coding process as illustrated in the previous chapter (see 5.1), therefore, the examiners' interview comments and the concepts extracted from the interview data (Tables 6-1 and 6.1.a) may provide a window to explore and understand the examiners' internal behaviours as well as various factors that may affect them.

Thus, for example, a general judgement approach emerged from the analysis of the interview comments. The conceptual property of an alternative perspective was identified in the Audience Point of View criterion (5.4.2). When an examiner cannot make a decision by one criterion, another criterion may be used (5.7.4). This approach of changing perspectives and criteria is observable assessment behaviour for many of the examiners. Comment 15 shows one such typical approach in choosing the better student interpreter.

Comment 15 (translation)[50]: […] I will pay more attention to see if there are

[50] **Comment 15** in source text Chinese: […]我會比較注意那個意思上的有沒有誤譯，[…]會很注意說，耶，這邊有沒翻錯。[…]現在有的情況的話…A 跟 E 哪一個比較好？A 跟 E…這麼一個小小的段落，然後…嗯，(long pause)…他們意思掌握上都各自有一些錯誤啦，然後聲音表情上的話，[…]所以主要原因是什麼呢？…我覺得這兩位同學很難取捨，[…]這兩個我大概選 E。（*為什麼選 E？*）能繼續聽下去吧…A 聽起來就是比較急一點。A 可能他好像等的時間比較久一點，然後等到他好像聽得差不多，他就很快很急地講出他記得的事情這樣子。

> meaning errors in the interpretation. […] I will check carefully to see if there is a mistake here or there. […] Under the circumstances…which one is better, A or E? A and E…in such a short paragraph, and then…hum…, *(long pause)*…well, they all had some meaning errors, and their voices and deliveries…[…] so what is the main reason?...I feel it is so difficult to choose between these two. […] I probably will choose E. (*Researcher asked: Why choose E?*) Well, I can keep listening…A sounded a bit rush. It seems that A waited longer to interpret, and then when she had listened enough, she blurted out very quickly what she remembered.

From Comment 15, it is clear that the examiner switched from Fidelity and Completeness to Presentation and Delivery as the decisive criterion to choose the winner. This happens when the accuracy levels of the two student interpreters are very close to each other, making it difficult for the examiner to choose; so, a second criterion is used to make the decision. Comment 16 is another example of this kind of approach to judgement. The examiner shifted from a predominantly fidelity-minded teacher point of view to the more user-friendly perspective as a member of the audience.

> **Comment 16 (translation)[51]:** I feel that it's very difficult to compare because it's just as I said, I emphasise accuracy. So, when both have problems with accuracy, I compare their delivery and presentation. I would consider the fluency of expression (the ideas), the words used and whether or not the audience can actually understand you. These are the things that I care about.

In other words, this examiner applied the assessment criteria in the order of fidelity, completeness, and then delivery. Therefore, we will refer to this as the **Fidelity-Completeness-Delivery (FCD) approach**. This approach is generally adopted by most of the examiners in this study, and conforms to the finding that in terms of their usage percentages (5.7.1), the primary criteria used by the examiners are Fidelity and Completeness and Presentation and Delivery.

[51] **Comment 16** in source text Chinese: 我覺得很難比因為就我剛剛已經提到，我比較重視的是 accuracy 嘛，那兩個人在 accuracy 上面都有問題的情況下，我可能再來要比的是他們在 delivery 跟 presentation 的部份。我就會考慮到說，你在表達的時候 fluency 啊，還有你表達的字眼啊，觀眾聽起來到底可不可以聽得懂，這是我比較在乎的。

Normally, a general approach that applies to most examiners would benefit the assessment procedure, making it more consistent and reliable. However, it was found that judgement fluctuations occurred among the examiners (4.1), that there are variations in the conceptual properties of the assessment criteria (5.2-5.6), and that the examiners used the criteria in different manners when assessing students (5.7). Given these findings, we need to ask why those variations occur and how or whether the FCD approach in practice contributes to the consistency of the examiners' judgements.

In the following sections, more internal assessment behaviours of the examiners will be discussed – examiner attention (6.2), examiner bias (6.3), and professionally-referenced standards (6.4).

6.2 Examiner attention

6.2.1 Different levels of attention to details

There are variations among the five clusters of examiners in terms of their attention distribution. Comparing the two super examiners that have the most members, PCC1 and PCC2, the latter seems to pay attention to a wider spectrum of issues in the students' simultaneous interpreting performances. They speculate more on the students' performances, especially on their interpreting strategies (also see 6.4).

However, there is a limitation to the examiners' attention span and memory load. For example, from the researcher's on-site observation and the comments made by the examiners in interview, PCC2 examiners in general appeared more often to miss or forget specific details, especially when the speech script was not used. Some examiners pay more attention to the frequency of mistakes and would simply say something like "there are more errors" or "fewer mistakes", but they were usually unable to give a specific account of the mistakes. Occasionally, a figure was given to indicate the amount of errors in a student's interpretation. For example, "there is a 10% difference between the two student interpreters". However, the 10% figure is almost certainly just a figure of speech to indicate a small difference, not because the examiner took notes of the errors, counted them and calculated the percentage difference.

The examiners in PCC3, PCC4, and PCC5 were frank in admitting their inadequate attention during the assessments, and often said, "I forgot", "I'm not sure", "did not hear clearly". PCC4 and PCC5 examiners also did not go into much detail when commenting on the Fidelity and Completeness criterion. They often made very general comments,

such as "almost all wrong", or "completely wrong" and did not comment on specific examples of mistakes. Comment 17 below is a typical sample of this kind of general comment on the Fidelity and Completeness criterion. The parts that relate to Fidelity and Completeness are underlined.

> **Comment 17 (translation)[52]:** I would think that B translated more content. The audience would receive more messages. Because now I have the script in front of me, I can tell in greater detail that A's listening comprehension could not keep up very well, missing key words in many sentences. B is pretty good, she is also very steady, […] Between A and B, it is obvious that B is better than A. The main reason is that firstly her sentences are cleaner and more complete. I believe that her listening comprehension is also better; she listened and got the key words. As for A, she probably just got some words here and there and tried to make them into sentences. And the meanings were very far away from the original.

In Comment 17, it is clear to see that the examiner considered many different criteria. However, in respect to the Fidelity and Completeness criterion, the examiner was vague and did not give specific details, such as which key words were missing, or what messages were "very far away from the original".

In contrast, Comment 18 contains much more details about the two students' performances. As the comment is very long, the parts that relate to the Fidelity and Completeness criterion are *partially* extracted here for illustration.

> **Comment 18 (translation)[53]:** (Ally) couldn't quite keep up when it comes to

[52] **Comment 17** in source text Chinese: 我會認為 B 他所翻譯的內容比較多，那聽的人有收到更多的訊息。那現在因為我面前有個稿子，我可以更仔細地看出來 A 的聽力稍微跟不上，有很多的句子裡面的關鍵字他都沒有抓到。那 B 是相當地優秀，而且也很沉穩，[…]在兩位之間是 B 明顯地比 A 優秀。[…]主要原因是第一點，他的句子比較完整乾淨，我相信是因為他也，他的聽力也比較好，就是他有聽懂了，而且有抓到關鍵的字。那 A 的話可能就是，片片斷斷地抓到幾個字，然後想辦法湊出句子來，然後那個意思跟原來的差得很遠。

[53] **Comment 18** in source text Chinese: (A)在數目字方面也比較有一些地方跟不上。有些關鍵字他沒有抓到，比如說一開始他講了一個很重要的重點是好比像一個交響樂團，他就沒有翻到。[…]數目字呢處理也不是很好，比方說公司的營業額是 11 億的，他翻成 12 億。[…]至於 B 的時候呢，他能夠把交響樂團翻得出來，[…]還有一個在 B 呢比較嚴重的錯誤，他因為這家公司介紹他是做地毯跟這些紡織品的時候呢，他講成變成地產公司。可能是心裡面想地毯，但是結果講成地產，這是常常

> numbers. She also didn't catch some key words. For example, in the beginning he (the speaker) mentioned an important point, using an orchestra as an analogy, and she (Ally) didn't get that part. […] The numbers weren't dealt with properly, either. For example, the company's business volume was 1.1 billion, she interpreted as 1.2 billion. […] As for B, she was able to interpret the orchestra part. […]B also made another more serious mistake. When the speaker was introducing the company as a carpet and fabrics manufacturer, she (Beth) interpreted it as a real estate company. She might be thinking *di tan* (carpet) but it came out as *di chan* (real estate), which is an easy and common mistake in simultaneous interpreting. […] Beth could make people understand in general terms, but she needs to improve on using business terms.

The two comments clearly illustrate the difference in the amount of detail given by the examiners. Of course, giving fewer details in their comments does not necessarily mean that the examiners did not pay attention to details. Some examiners, for example, just made notes of errors on the scripts whereas some examiners, perhaps with better memory or experience, were able to give more detailed comments. It may also be that some examiners relied on their overall impressions to judge students (see also 6.2.2) while some others have a more systematic assessment approach, such as relying on note taking and using the speech script.

How, then, do these differing levels of attention to detail affect the examiners' decision-making? Let us use the detail *di-tan* in Comment 18 to scrutinise and answer the question. Beth interpreted the speaker's company as a *di-chan gongsi* (real estate company), whereas she should have said a *di-tan gongsi* (floor carpet company). In the Daisy-Beth comparison, most PCC1 examiners either did not mention this or did not pay attention to this detail. Nine of the ten PCC1 examiners in this cluster picked the same winner Beth. Only one examiner did not pick Beth, saying that her getting the company product wrong is a very serious mistake.

有的做同步翻譯時候的通病。[…]B 這方面呢他整體翻的大家明白，但是就是在商務用詞方面還需要改進。

However, this examiner acknowledged that she did not notice that Daisy had made the same detailed error, and made the following comment, which is a clear acknowledgement of an examiner's limited attention span to details.

> **Comment 19 (translation)[54]:** Regarding this (mistake) in D's interpretation, I didn't actually notice. She might have also made the same mistake and I just didn't catch it.

This examiner is an interpreter examiner in PCC1, which consisted of mostly non-interpreters (Table 4-9.a). One thing to consider is that even when this examiner knew that she did not pay adequate attention to Daisy's interpretation and might have missed a mistake she had noted with another student, she, nonetheless, made a decision based on incomplete information.

It may not be reasonable to expect an examiner not to miss any details. However, when the missing detail is the key element for making important judgement in an examination, such as in the case of Comment 19, the results of the examination would not be very reliable. Therefore, this seems to be more than a problem of lacking attention to detail, but an assessment behaviour that could be improved so important details are not missed, for example, by examiner training and support for using proper assessment tools to ensure a more reliable examination procedure.

Another example to illustrate the effect of attention span on the judgements is PCC2's Beth-Cherry comparison (also see 5.7.4). In this paired comparison, the examiners made different choices of winner. One of the examiners quickly made up her mind that Cherry's interpretation was full of mistakes. As the examiner did not use the speech script when listening to the student interpreters, the researcher asked the

[54] **Comment 19** in source text Chinese: 關於這個 D 這邊我並沒有注意到，她可能也弄錯了，只是我沒有抓到而已。

examiner to clarify by giving examples of Cherry's mistakes. Comment 20 below was the examiner's response.

> **Comment 20 (translation)**[55]**:** Could you give me the speech script? I forgot what they were. *(after reviewing the script)* […] I have forgotten what her wording was, but I know it was wrong there.

This kind of response may cause problems at an examination panel when deliberating on a decision on student interpreters' performances. It is not very convincing for an examiner to make a judgement but be unable to give supporting evidence, especially when there are contradictory views. If examiners did not pay consistent attention to student performances, it would be unlikely that reliable judgements could be made. It is clear that the factor of examiner's attention plays an important role during the judgement process. The uneven level of attention among the examiners, therefore, might well be the reasons underlying the inconsistencies of assessment results.

6.2.2 Judgement by impression

From the previous examples and discussions, we have seen that some examiners did not pay enough attention to details. This could mean that those examiners might have made judgements based on their impressions of the students' performances. For example, when comparing Cherry and Eileen, an examiner simply said,

> **Comment 21 (translation)**[56]**:** "I didn't take many notes about C and E. I *felt* that E is better", (*emphasis added*).

[55] **Comment 20** in source text Chinese: 你可不可以給我稿子，我忘記是什麼了。[…]我已經忘記他用的 wording 是什麼，可是我知道那裡是錯誤的。

[56] **Comment 21** in source text Chinese: C 跟 E 我沒有記下太多筆記，我*覺得* E 比較好。

This examiner had only written a few words for each other student, but *none* for Eileen. Therefore, this examiner must have made her judgement based on her general impression, i.e. "I felt that Eileen is better." Interestingly, this examiner was the only one in PCC5 that picked Eileen (see Appendix D). Given that the decision could only have been based on the examiner's general impression, could the judgement be considered sound and accountable?

Comments 22-24 illustrate how two PCC4 and PCC5 interpreter examiners made their judgement by impression. The examiners were mainly commenting on Ally's interpreting performance in the Eileen-Ally comparison. In this pair, one examiner said that Ally's interpretation was *totally wrong* (全錯) and not acceptable, and that Eileen's delivery was annoying but, in terms of faithfulness, not as dangerous as Ally's. Nonetheless, when asked to give one or two examples of Ally's serious mistakes, the examiner responded with the following remark:

> **Comment 22 (translation)[57]:** I didn't write them down, but because…sorry, I didn't hear very clearly why (they were wrong) because if she (Ally) was wrong from the beginning to the end in the process of the examination, I wouldn't bother to remember the details. Because basically she is a definite fail […] not borderline, I won't spend more time to check. It's just wrong, wrong, wrong….

This examiner clearly did not pay careful attention to Ally's interpretation and reached a quick judgement, i.e. totally wrong so a definite fail. Ally was the first student to be assessed. After this first impression (also see 6.3.1), the examiner did not attend carefully to the rest of Ally's interpretation.

Comment 23 (translation)[58]: (Ally's) whole interpretation is a blur. […]

[57] **Comment 22** in source text Chinese: 剛才我沒有寫下來，但是因為…對不起，我沒有聽得那麼清楚是為什麼，因為在考試的過程中如果她從頭錯到尾的話，我就不會再去記得更細了，因為基本上她是一個 definite fail，[…]不是 borderline，所以我不會花更多的時間去看，就是錯錯錯…。

[58] **Comment 23** in source text Chinese: [A's]整段全篇的翻譯幾乎是全部一團模糊。[…] 基本上來

> Basically, after listening to her whole interpretation, I don't think that the audience would be able to understand what she was saying. I could not understand the messages she was presenting, even though they were in Chinese.

Comment 23 also reveals a strong personal impression; this examiner felt that Ally's interpretation was beyond comprehension. It seems to be an even worse impression than the 'totally wrong" impression of Ally's interpretation in Comment 22.

Nevertheless, in contrast to the above negative impressions, a non-interpreter PCC4 examiner, seems to have been able to see Ally's interpreting performance in a more balanced way. The examiner first commented that neither Ally nor Eileen performed well, but Ally's accuracy was slightly better than Eileen's. The examiner considered the choice and finally picked Ally, making the comment below.

> **Comment 24 (translation)**[59]**:** If purely (judging) as a listener, I probably would feel that E is more comfortable to listen to. But today I am not a listener, I'm examining, I will choose A because I emphasise accuracy more. *(Researcher asked: In a school exam?)* Yes, I will choose A. (*Researcher asked: But what if this examination were to select an interpreter for a job?)* For selecting an interpreter for work, I think I would still choose A, […] (because) accuracy is more important.

Holistic impressions may not necessarily be inaccurate, but they lack details and therefore do not allow for nuance. Different examiners may pick up different details based on their different perceptions and attention span. In addition to the above examiner that made Comment 24, for example, all of the PCC3 examiners also picked Ally (Table 4-10), saying that her fidelity was higher than Eileen's. These examiners who picked Ally must have been paying attention to something that the other examiners

---

說，全篇聽下來，我覺得觀眾不可能聽懂她在講什麼東西。我聽不懂她在講什麼訊息，雖然她講的是中文。

[59] **Comment 24** in source text Chinese: 如果是純粹聽眾的話，我可能會覺得 E 比較聽得舒服一點。如果我今天不是一個聽眾，我要考試的話，我會選 A，因為我比較重視正確性。（*在學校裡面的考試？*）對，我會選 A。（*但是如果這個考試是為了要選口譯員去工作呢？*）選口譯員去工作的話，我看我還是會選 A 耶，[…]正確還是會比較重要一點。

had not been able to see in Ally's interpreting performance, and they all made the decision based on the fidelity criterion like those who did not pick Ally as the winner. Again, here we see a familiar situation where different examiners use the same criterion, i.e. fidelity in this case, but made contradictory judgements (5.7.3).

What could have affected the different perceptions of the interpreter examiners, such as the ones in Comments 22-24 above? What factors could have limited some of the examiners' understanding of Ally's interpretation so they had such a negative impression on her interpreting performance? Those examiners clearly relied more on their impressions, especially when a very quick judgement was made, saying that Ally's interpretation was a blur and all wrong. Here, the lack of attention may have been caused by a strong factor, the first impression effect, or a bias in the examiners' assessment behaviour.

6.3 Examiner bias

From Tables 6-1 and 6-1.a (pp. 213-214), examiners' biases were also found, such as the primacy-recency effect (6.3.1) and personal preferences (6.3.2). This section presents and discusses these biases by using more sample comments for illustration.

6.3.1 Primacy-recency effect

An interesting conceptual property regarding examiner behaviour is the examiners' awareness of the *primacy-recency effect*. In psychology and sociology, the primacy-recency effect is a cognitive bias. To put it simply, a primacy effect refers to the greater impact of what we first learn about someone, i.e. the first impression; a recency effect happens when the later impression predominates (Luchins, 1957). After examining the five student interpreters, examiners commented that the order in which they observed the student performances may have influenced their perceptions of the students' interpreting abilities. If the first student performs very poorly, extra credit may be given to the later ones even when in reality their performances may not be significantly better than, or in some cases, not as good as the student giving the first impression. Comment 25 below illustrates a typical comment relating to such a view.

> **Comment 25 (translation)[60]:** In fact when I listened for the second time, I had some doubts about my previous judgements. The notes that I had made previously were more of a general impression, which I feel had some "anaesthetic" effect.

[60] **Comment 25** in source text Chinese: 事實上是我第二次聽的時候對我之前做的會有懷疑，就是我之前寫的 notes 可能是一個比較 general 的 impression，而且我覺得這事實上有一點點"麻痺"的效果，就是做得都不是特別好。而且那個順序也有差，如果說第一個人做的是特差，第二個人第三個人做的雖然也不好，你就會覺得還不錯，可以接受。[…] 如果第一個人做的很不好的話，對於後面的人來講是加分的效果。

> The students did not interpret particularly well, and the sequence of listening to them made some difference [in judgements]. If the first one is very poor, you will then feel that the second and third ones are not bad and acceptable, even though they may not do well, either. […] If the first one did not do a good job, there will be a tendency to give more marks to the later ones.

Theoretically speaking, this primacy-recency effect is likely to happen to most, if not all, examiners and comments about this effect can be found in every cluster. Those examiners who reviewed the recordings are more likely to notice this effect. This effect may also explain why more than two thirds of the examiners did not pick Ally, with her nervous delivery, as the winner. Ally was the first to be assessed, followed by Beth whose delivery was regarded as calm and confident by many examiners.

Therefore, this examiner bias can be powerful in affecting the way examiners make judgements. The primacy-recency effect coupled with the impressionistic judgement mentioned earlier may create a structural problem of the interpreting examination, in which the order of student interpreters being assessed may affect the way an examiner perceive their performances. For example, as Comment 22 has illustrated (6.2.2), the negative impression on Ally was so strong that the examiner did not listen to her performance thoroughly and made an immediate judgement that she had failed.

6.3.2 Personal preferences regarding interpreting and delivery styles

Examiners' personal preferences in relation to delivery style and interpreting in general were found in Table 6-1. Just like the primacy-recency effect, personal preference could also be a powerful source of influence on the way examiners make decisions. In PCC1, for example, examiner R28 had the most disagreements with the other examiners (Table 4-10). R28 chose Ally in three pairs – Ally-Beth, Eileen-Ally and Daisy-Ally, while *none* of the other examiners picked Ally in those pairs. R28 made

it clear that her personal preference regarding an interpreter's delivery would influence her choice of interpreters. She was the only examiner in this cluster whose decisions were based mostly on the delivery criterion rather than on fidelity. This is a case that clearly shows how an examiner's preference influences the decisions made.

Another example is the Eileen-Ally comparison in PCC2. In this comparison, most examiners focused on the students' delivery and strategies, which according to the examiners resulted in interpretations with different levels of fidelity and completeness. Eight out of the ten PCC2 examiners chose Eileen as the winner (Appendix D). Comment 26 below is a comment typical of their views.

> **Comment 26 (translation)[61]:** Overall speaking, both (students) had a lot of mistakes, but I like Eileen's interpretation better because I feel that Eileen was more fluent, not in such a hurry. […] In this sense, therefore, I think Eileen is the better one.

The eight examiners considered Eileen to be more fluent than Ally, which is in contrast to the judgements of the other two examiners who chose Ally as the winner. One of the two examiners commented that she disliked Eileen's delivery even more, though Ally's delivery was fast and sounded a bit nervous. This examiner said,

> **Comment 27 (translation)[62]:** As for E, I cannot stand listening to her. She kept saying "wuo men, wuo men, wuo men (we we we)." Very jerky delivery, and her sentences were not very complete. It's uncomfortable when listening to her, when listening to her for a longer time it may be uncomfortable. I will still choose A.

That is to say, this examiner would rather listen to Ally's nervous delivery than listen to Eileen's jerky interpretation. It is clear that the PCC2 examiners had preferences for the

[61] **Comment 26** in source text Chinese: 整體來講的話，雖然錯誤兩個都蠻多的，但是我會比較喜歡 E 的翻譯。因為 E 的翻譯我覺得比較流暢，比較沒有那麼急促，[…]所以就這方面來講的話，我覺得 E 會比較好。

[62] **Comment 27** in source text Chinese: E 的話我看，我很受不了她的「我們我們我們」，很 jerky，就是她的一個句子沒有辦法很完整。聽起來蠻不舒服，聽久了可能蠻不舒服。我還是會選 A。

interpreter's delivery style, which played a part in making their judgements. For the examiner making Comment 27, the recency effect might also have played a part in the judgement process because that examiner said repeatedly in later paired comparisons that she did not like Eileen's jerky delivery.

Another examiner in PCC4 also did not like Eileen's delivery style. The examiners' personal preference was so strong that it was enough to influence the examiner to deviate from the FCD approach when comparing Daisy and Eileen. The examiner felt that Eileen's delivery was very annoying and made the following comment.

> **Comment 28 (translation)[63]:** E's delivery is horrible. It needs to be greatly improved. […] Although she managed to make a lot of points, toward the end I couldn't stand listening to her. […] This kind of up and down, this kind of intonation is very tiresome to the audience.

Comment 28 shows that even though this examiner knew that Eileen "managed to make a lot of points", she still would not pick Eileen because she could not stand listening to Eileen's delivery style. Because of her personal preference, this examiner abandoned the FCD approach and based her decision on Presentation and Delivery rather than on the Fidelity and Completeness criterion. The examiner also said that Eileen sounded childish, but interestingly, a PCC1 interpreter examiner commented that Beth, rather than Eileen, sounded childish and too girly.

From the contrasting views above, we can see that in terms of delivery, while many examiners may disfavour a nervous delivery, some examiners may have stronger reactions to certain delivery styles of the interpreter. This factor of personal preferences does play a role in influencing the examiners' decision-makings.

---

[63] **Comment 28** in source text: E's delivery is horrible. It needs to be greatly improved. […] 雖然很多 points 都說出來，可是到最後我已經聽不下去了。[…]這種 up and down 的話，這種 intonation 對於觀眾來講是很累的。

The FCD approach and personal preferences

Nevertheless, if the personal preference is not too strong, the examiner is still likely to follow the FCD approach. Take PCC4's Beth-Cherry comparison for example. PCC4 consists of only two examiners, and both examiners agreed that Cherry made fewer errors and chose her as the winner. However, one examiner thought that Cherry's delivery was *sloppy* and not concise enough[64] and the other examiner commented that Cherry's delivery was *steady* and could keep up with the speaker[65]. In this example, the two PCC4 examiners had very different preferences for the style of the interpreter's delivery, but both examiners picked the same winner Cherry based mainly on the same criterion of fidelity.

The fact that the two examiners chose the same winner indicates that the weighting of the Presentation and Delivery criterion is less than that of the Fidelity and Completeness criterion. In other words, when a personal preference is not too strong, the examiners still follow the FCD approach. The Fidelity and Completeness criterion will be the first to be considered when making a decision, and only when the primary criterion cannot separate the student interpreters' performances will Presentation and Delivery be used as the decisive criterion.

[64] Examiner One's comment in Chinese: 從語言來講，C 很 sloppy。就是他懂得這個意思，但是他沒有辦法很清楚很簡潔地表達出來，他不夠 concise。

[65] Examiner Two's comment in Chinese: B 的誤譯的地方就比 C 多。C 跟得上原來 speaker 的速度，然後整體的速度也很平穩，所以我覺得 C 比 B 要好。

6.4 Professionally-referenced standards

For reasons of test validation, performance assessment often relies heavily on professional judgement of the subject area concerned (2.3.2.b). Here, the conceptual property of professionally-referenced standards was identified from within the comments made by the examiners, especially by those who have a background in interpreting. Interpreter examiners can compare the student interpreters' performances with that of an experienced interpreter, or make comments about how things should be done in the real world to achieve the intended communicative effect, and what the audience would think of the student interpreters' performances.

This concept of professionally-referenced standards relates not only to the implementation of the Fidelity and Completeness criterion, but also to how different criteria are weighed against one another when assessing the interpretation. For example, PCC2 differentiates the Fidelity and Completeness criterion into different levels according to different situations; PCC3 believes that "accuracy covers rushed delivery", which means that poor delivery is acceptable as long as it achieves message accuracy; PCC4 tolerates omissions more than errors, and PCC5 holds a similar view: being vague is worse than having omissions, especially when an interpreter's comprehension is problematic.

These conceptual properties are the examiners' subjective judgements that may have been formed and based on their professional experiences. In other words, to a certain extent, interpreter examiners are referring to their professional standards when assessing the student interpreters. In the sections below, more sample comments will be used to illustrate and discuss these assessment behaviours.

6.4.1 Situational weightings of assessment criteria

Previously in Chapter 5, it was found that examiners differentiated the fidelity criterion along different dimensions (5.3.2). This may be in part related to the multi-dimensional perspectives on quality in interpreting (2.1.2). The interests and motivations of the different parties to an act of communication in a specific situation need to be considered in a quality assurance scheme for interpreting (2.1.5). This multi-situational consideration also appears in interpreting examinations, which helps validate the examiners' assessment approach. Comment 29 below illustrates the general view of an interpreter teacher on the different focuses of the interpreting examination.

> **Comment 29 (translation)[66]:** […] The direction of assessment in each interpreting exam…some teachers may have different focuses. For example, studying in different [interpreting] courses, certain criterion or assessment standard may be particularly emphasised at certain stage of learning.

Since the source speech of the examination task in this study was set in a business conference, many examiners weighted the assessment criteria accordingly to a business situation. For example, many examiners regarded certain details in this business speech, such as numbers, monetary unit, the company, i.e. a *di-tan gongsi* (floor carpet company), and the use of business terms like "turn over", "market share" as very important. One examiner in PCC3 felt that it was very serious for an interpreter (Beth) to get the company product wrong (also see Comment 18 in 6.2.1). This error was weighted so much that the examiner sacrificed the internal consistency of his own overall judgement results and did not pick Beth as the winner (also see 4.5.2).

[66] **Comment 29** in source text Chinese: […] 每一個口譯的評鑑的方向…有些老師的重點不太一樣，比如為了學習不同的課程，學習不同的階段可能會在某一個階段特別注重某一個 criterion，某一個評鑑標準。

In respect to the Presentation and Delivery criterion, some interpreter examiners in this study, especially those who are active in the market without teaching duties, also held an interesting view that an interpreter should "keep talking", and thus could get more credit when interpreting in a business situation; an economical mode of delivery may not always be preferred. Comment 30 below illustrates this point of view.

> **Comment 30 (translation)[67]:** In my personal and subjective opinion, if I were a client who didn't know English and just came here for business purposes, I would like A better because she speaks more. This is a feeling in the market. C gives a concise delivery, but if it is too concise, the message or information in her delivery becomes just a skeleton and is not enough. Therefore, a client would feel that A gives a more detailed interpretation, whereas in the eyes of a person who doesn't know simultaneous interpreting, it would seem that C doesn't know how to translate, but actually C is only translating in a more concise way.

This examiner's view of "keep talking is better" may be different from many interpreter teachers who train their students *not* to be wordy, but to be as concise as possible. Most of those examiners that chose Cherry over Ally indicated that they liked Cherry's clear and concise delivery. The contrasting views, of course, may be in part due to personal preferences, but in the case in question here the examiner in Comment 30 made the judgement based on his experience in the field and emphasised the delivery criterion in his assessment, i.e. a professional judgement.

This market-oriented assessment approach was also noticed in some other interpreter examiners. The Audience Point of View criterion requires fluent and clear delivery so that the audience will keep listening to the interpretation (5.4.1). One PCC5 examiner takes this market- or audience-oriented approach to an extreme in assessing

[67] **Comment 30** in source text Chinese: 以我個人主觀來講，我如果聽不懂英文，我是普通的一個來做生意的來賓呢，我會發覺我會比較喜歡 A，因為好像他講話比較多一些。這是一個市場上的感覺。那 C 呢他雖然是講得很精簡，但是有時候精簡過頭了，變成 message 或是它裡邊的 information 就不夠，所以業者來講，他會覺得為什麼 A 翻的東西好像比較詳細，那以不懂同步翻譯的人來講呢，他會覺得 C 好像不懂得翻，而其實 C 只是用一種簡短的方式翻。

interpreting. Based on the market experience, this examiner emphasised the Presentation and Delivery criterion and followed this principle throughout the ten paired comparisons. In the Daisy-Ally comparison, however, a dilemma occurred. The examiner felt that Daisy's delivery was steadier but Ally's fidelity rate seemed higher. The examiner finally chose Daisy and explained the decision as in Comment 31.

> **Comment 31 (translation)**[68]**:** To me personally, […] delivery is more important than accuracy. So for these two [students] I will chose D as being better than A. […] You may feel that some things are important, such as accuracy or numbers, etc., but to the audience, they are not that important. I feel that the important thing is the feeling of the audience when they listen to your delivery [of interpretation], and to see if they can trust your voice.

Simply put, this examiner believes that good delivery comes first because it engenders a sense of trust in the audience. So a convincing delivery with some mistakes is still better than an accurate interpretation that is delivered in a hesitant way. This comment echoes Schjoldager's views that an interpreter's other qualities are irrelevant if the audience get irritated and lose interest in listening to the interpretation (Schjoldager, 1995).

The question is: can an examiner find a balanced and consistent way of weighing the two criteria, fidelity and delivery, which takes account of the situational variables of the source speech? In Comment 30, for example, the examiner also pointed out that Cherry used a more concise interpretation, but the audience might not be aware of the strategies involved. This also relates to the consideration of other assessment criteria, i.e. Interpreting Skills and Strategies, which seems to further complicate the matter. How do the interpreter examiners weigh the various criteria when assessing student interpreters? Comment 32 below provides an interesting insight to this question.

---

[68] **Comment 31** in source text Chinese: 就我個人來說，[…] delivery 重於正確性。所以這兩個的話我會選 D 比 A 好。[…] 有些你覺得很重要的事情，譬如說正確度或是數字什麼的，對聽眾來說並不是那麼重要，我覺得重要的是聽眾聽到你 delivery 的感覺，是不是能夠去信任你的聲音。

> **Comment 32 (translation)[69]:** The reason why I choose D is because although her expression is also poor, she knew how to say something neutral, which will not create big problems. […] As for A, she would quickly jump into a conclusion. […] I probably will choose D because overall speaking she caused less damage.

In other words, this examiner's decision was based on the consideration of damage control, which could be attributed to the consideration of interpreting strategy, i.e. a strategy to minimize the interference of information recovery (Gile, 1995a: 202).

These considerations about different weightings being given to certain types of messages, delivery styles and strategies were based on the examiners' professional experience of in the profession. Depending on the nature of the occasion, certain information or messages are deemed more important and must be delivered correctly in an acceptable manner.

6.4.2 Omissions vs. Errors

Juggling judgements between omissions and errors in student interpreters' performances is common for examiners when they are applying the Fidelity and Completeness criterion in consideration of the Interpreting Skills and Strategies. To cope with the cognitive overload in simultaneous interpreting, interpreters often have to operate on what Shlesinger called the "condensation norm" that

> "not only condones but often encourages strategic macroprocessing", so that "not every element of every proposition in the source text needs to be reproduced as such. It is appropriate for a simultaneous interpreter to produce the underlying meaning of the proposition" (Shlesinger 1999: 69 in Marzocchi, 2005: 92).

[69] **Comment 31** in source text Chinese: 我之所以會選 D 是的原因是因為她在 expression 上面雖然說一樣地不好，但是她懂得用稍微 neutral 一點的東西，比較說不會造成很大的問題。[…]那 A 很快地就 jump into conclusion. […] 我大概會選 D 吧，因為整體上來說，她造成的 damage 比較少。

Gile also argued that, "not all the information which was omitted in the target-language speech is necessarily lost as far as the delegates are concerned, since it may appear elsewhere or be known to the delegates anyway" (1995: 200). Shlesinger proposed the condensation norm on the basis of her literature reviews of interpreting studies; it has been intuitively corroborated by many interpreter trainers' experiences and is in line with the long-standing discourse on conference interpreting (Marzocchi, 2005: 92). As evidential support and for analysis, Comments 33-35 below illustrate how the examiners in this study applied this condensation norm when assessing student interpreters.

> **Comment 33 (translation)[70]:** D is worse than C. First, she (D) is not fluent enough; second, she omitted more messages, […]. Compared with C, however, because she (D) omitted a lot, there seems to be less error (in D's delivery).

The underlined part of Comment 33 illustrates the examiner's view of the relationship between omissions and errors when interpreting. In Comment 34, when comparing Daisy and Eileen, the same examiner further elaborated on which is the more serious – omission or error. The examiner concluded that overall Daisy was better.

> **Comment 34 (translation)[71]:** It's because that although she (Daisy) lost a lot of material, at least she did not make so many mistakes. I would rather see her omit things than see her say something wrong.

In other words, errors are less condonable than omissions. Surprisingly, this view seems to be shared by both interpreter and non-interpreter examiners alike, as similar comments were made by examiners from both backgrounds. Comment 35 sums up this assessment approach in weighing omissions against errors.

---

[70] **Comment 33** in source text Chinese: D 不如 C，第一個就是他不夠流利，第二個就是他遺漏的東西比較多，[…]可是跟 C 比較起來，因為他漏掉很多，所以好像錯誤的地方比較少一點。

[71] **Comment 34** in source text Chinese: 因為他(Daisy)雖然丟掉很多東西，至少他沒有犯那麼多的錯誤。我情願他漏掉東西，不要講錯。

Comment 35 (translation)[72]**:** I often feel that the most basic problem to consider in interpreting is: [we] would rather have omissions than errors in interpretation.

This omission tactic, however, should only "refer to the case where an interpreter *deliberately* decides not to reformulate a piece of information in the target-language speech" (Gile, 1995a: 200). In this study, some examiners also made a distinction between *not understanding* the message and *not hearing* the message at all (5.6.1). This was when applying the two criteria of Interpreting Skills and Strategies and the Foundation Abilities for Interpreting. Safe implementation of the omission strategy can only be achieved when the interpreter fully understand the messages and has the capacity to process them, i.e. to deliberately decide what and when to omit.

6.4.3 Interpreting skills and strategies

As mentioned earlier in Chapter 5, successful or poor implementation of interpreting skills and strategies can be observed by looking at the EVS in relation to the interpretation output (5.5.3). Taking the omission tactic as an example, observable evidence of this tactic in action is usually a longer EVS (because the interpreter needs to listen for longer to decide what and how to omit material) and a reduced or condensed amount of information in the interpretation output.

Depending on the professional experience of an interpreter examiner, however, there may be different ideas about when an EVS is long and how much reduction is acceptable when applying the condensation strategy. This subjective variation is probably why the Interpreting Skills and Strategies criterion is difficult to use, and therefore less favoured as a primary criterion. Nonetheless, to some extent it still affects

[72] **Comment 35** in source text Chinese: 我常常覺得說口譯可能最基本的問題應該還是，即使漏譯也不要誤譯。

the way an examiner makes the judgement (5.7.1), especially when multiple criteria are linked together in the judgement process.

Some contrasting judgement results in relation to Beth and Cherry have been observed and analysed in 5.7.4. Comments 36-40 below further illustrate different perspectives on how different examiners judged Cherry's interpreting strategies. First, Comments 36-38 are positive about Cherry's interpreting strategies.

> **Comment 36 (translation)[73]:** C follows more closely when interpreting. Her translated sentences are more concise and shorter, combining two or three sentences into one to translate. Therefore, comparatively speaking, I think she sounds more fluent from the audience point of view. […] I feel that C did a better job in reduction (strategy). […] Whereas B, she followed (the speaker) more slowly (with longer lags) so she often would easily miss…miss something.

> **Comment 37 (translation)[74]:** Overall C did better than B. The reason is that her judgment is more accurate than B in terms of deciding what messages need to be translated, especially when there are parts that have to be thrown away under time constraint. In other words, she would omit information that was not important. […] So overall just from an audience point of view, […] my personal feeling is that I know what the speaker was trying to say via the interpreting service.

> **Comment 38 (translation)[75]:** She (Cherry) would consider the coherency issue. […] Thus it sounded more fluent and coherent [in Chinese]. I feel that C has the ability of not translating in a rigid way. B could interpret correctly, C could also

[73] **Comment 36** in source text Chinese: C 他在翻譯的時候跟得比較緊，然後他在翻譯句子上也會比較簡短，就是他比較會把兩三個句子合併成一個句子來翻譯。這樣比較上來講，以觀眾的角度來講，聽起來比較順，[…]我覺得就是在做刪減上[…] C 比較好。然後因為 B 他在跟的時候跟得比較慢，所以他常常比較容易會漏掉…漏掉東西。

[74] **Comment 37** in source text Chinese: 總體上來講 C 比 B 做得好的原因是，他對什麼地方應該要翻出來，或實在來不及要丟掉的地方，在他的判斷上面比 B 要來得準確一點。換句話說，不是非常重要的 information 他可以丟掉它 […]。所以整體上如果是純粹的聽眾，[…] 我個人覺得 C 給我的感覺是，我知道原來的說話人要說什麼，透過口譯的服務。

[75] **Comment 38** in source text Chinese: 他會考慮到首尾連貫的問題。[…]這樣子聽起來會比較通順，我覺得他好像有這樣子的能力。就是不是完全這樣子硬翻。那 B 也會翻對，C 也會翻對，但是 C 顯然多做了一個潤飾的工作。[…] 對聽眾來講，我也不用花太多時間腦筋去思考，我就是聽你在演講，一個…比較像一個中文的演講。

> interpret correctly, but C obviously did an additional polishing job. […] As a listener, I don't need to make much effort to think, I just listen to a speech [via interpretation], a speech…a speech that sounds more like in Chinese.

As we can see, the above three comments show high regard for Cherry's reduction strategy in her interpreting performance, and her concise and polished delivery in the target language, as well as a good level of fidelity in the messages. The final decision was then based on the overall quality determined by the interpreting skills demonstrated and considered from the Audience Point of View. These criteria are closely linked.

However, there are also examiners who thought differently about Cherry's interpreting performance, such as Comment 20 in 6.2.1 and some PCC2 examiners' negative comments on Cherry in 5.7.4. Comments 39 and 40 below further illustrate the different observations on Cherry's interpreting skills.

> **Comment 39 (translation)[76]:** I like B better. […] I feel that she might have understood the messages, but her Chinese was not well delivered in terms of word choice. […] She has a strange way in Chinese delivery. In terms of content, there are also some mistakes. She had more obvious mistakes than B.

> **Comment 40 (translation)[77]:** I feel that B is better. This is because overall B was very steady. […] Sometimes C just chased up a few words in the sentences, and then used them to make up her sentences, but the meanings could be far away from the source text. Sometimes the gist was not even there. […] very fragmented like bits and pieces of a jigsaw. […] Yeah, I almost certainly did not hear any specific facts [from C's interpretation].

[76] **Comment 39** in source text Chinese: 我會比較喜歡 B，[…]我覺得他可能聽懂，可是他的中文處理得不夠好，選詞用字上面，[…]有一些比較奇怪的處理方式。那內容來說的話，內容也是有一些錯誤，內容的錯誤是比 B 的更明顯的。

[77] **Comment 40** in source text Chinese: 我覺得 B 比較優秀，原因是 B 整體來說非常地沉穩，[…] 那 C 有時候是追到句子裡面聽到幾個字，然後就拿那幾個字來湊句子，跟原文的意思可能相差了很遠，甚至呢主旨都沒有表達出來。[…]非常地零星，零碎，所以真的是拼湊出來的圖。[…] 對，幾乎沒有聽到任何具體的事實。

These two comments are very different from Comments 36-38. Comment 39 talked about Cherry's "strange" Chinese delivery. Given that many other examiners liked Cherry's concise delivery, this could just be a matter of the examiner's preferences in regard to the interpreter's delivery style. However, the comment on Cherry's "obvious" mistakes implies that there are significant problems underlying the examiners' different views. Cherry's interpretation was considered as "concise and polished", but in Comment 40 it became "fragmented like bits and pieces of a jigsaw", which impeded the examiner from obtaining the speaker's messages via the interpretation.

Could the contrasting views be attributed simply to the factor of the examiners' preference bias in regard to delivery styles as discussed in 6.3.2? Or could the preference bias extend to the implementation of interpreting skills and strategies? Comments 41 and 42 below may give some clues to answer this question.

> **Comment 41 (translation)[78]:** She (Beth) actually hasn't got big mistakes, just lagging too far behind the source speech. I have a deep impression that after she had listened to some segments of the speech, she would keep talking and explaining. She would only go back to listen to what the speaker was saying after she finished explaining (a segment). She was not really doing simultaneous interpreting. I feel that it's summarisation with some explanatory translation. […] She omitted a lot. It's just that she was very lucky because the omitted parts were by chance not the important ones. It's true. This is my opinion.

In Comment 41, the examiner was comparing Beth and Cherry, and picked Beth as the winner. Compared to Comment 36, however, the examiner in Comment 41 seems less critical of Beth. The examiner picked Beth even though she had said that Beth was not really doing simultaneous interpreting and omitted a lot of information, saying that she was lucky the messages she had omitted were not important.

[78] **Comment 41** in source text Chinese: 他(Beth)其實沒有什麼很大的錯誤，只是講的跟原文的間隔拖太遠了。我印象很深刻的就是，他會聽到某一個段落然後就自己講自己的話，一直解釋，然後等他的話解釋完才會去聽講者講什麼，而不是真的在做同步，給我感覺他是 summary 再加上解釋性的翻譯。[…]他是漏蠻多的，只是運氣非常好，漏的剛好都是不重要的。真的，這是我的看法。

In contrast, the examiners in Comments 36-38 believed that Beth's lags were long and resulted in the frequent loss of information, and instead expressed more approval of Cherry's reduction strategy and concise delivery. One possible explanation of these different views on the interpreting strategies used by the same students is that the examiners have their own preferred practices or strategies in interpreting. In other words, based on their professional habits, such as their preferred strategies and skills for interpreting, the examiners were *guessing* how the student interpreters did the job and judged the students accordingly.

Comment 42 below illustrates this professionally-preferred practice, which was made by the same examiner as Comment 41. In simultaneous interpreting, this examiner said that her preferred practice is to segment the sentences and follow the source speech with short lags, making it easier for her to do multi-tasking. This examiner made the following remark about Ally's multi-tasking skill:

> **Comment 42 (translation)**[79]**:** You will see that A always started to interpret when the speaker paused, and she usually spoke at the same time with the speaker. I couldn't hear what the speaker was saying at all. If I couldn't hear what the speaker was saying when listening to both languages at the same time…if I couldn't hear the speaker, how could she have heard?

The last remark in the examiner's comment is an interesting point, and it shows that this examiner was clearly using her own preferred practice to judge Ally's interpreting strategy. In other words, the personal preference bias may extend to the implementation of the Interpreting Skills and Strategies criterion (also see 6.4.4).

It is natural that interpreter examiners have preferred professional practices. As mentioned before, professional judgement is an important element in maintaining test

[79] **Comment 42** in source text Chinese: 你會發現 A 在講話的時候，都是利用講者的空檔，然後常常他講話的時候跟講者是同時進行，我根本聽不到講者在講什麼。那如果我用雙語兩個一起聽都聽不到講者在講什麼，他怎麼可能聽得到講者在講什麼。

validity and reliability in an interpreting examination. Interpreter examiners share similar experiences and expectations of the standard for interpreting; so they should be able to make relatively consistent judgements. Nevertheless, differences exist among individual examiners in terms of their interpreting experiences in the market place; they may also have different personal preferences. When such individual examiners are placed on the same examination panel, different opinions may occur.

The question is: to what extent will this professional or personal preference in relation to interpreting skills and strategies affect the actual final decision? From the comments above, we saw that the examiners might choose different winners, for example in the Beth-Cherry comparison, but their reported main criteria were still the primary two: fidelity and delivery. The Interpreting Skills and Strategies criterion seems to be more like a hook that links the two main criteria, and serves to explain how the quality of the student interpreters' interpretation was achieved, or affected.

This indicates that the Interpreting Skills and Strategies criterion may be more useful in formative assessment for giving feedbacks than for outcome-based judgement in summative assessment. Understanding how the examiners link these assessment criteria also helps us understand the relations between the criteria and how they may be made operational for the interpreting examinations, i.e. how the constructs of interpreting are measured.

6.4.4 Diagnosing student interpreters' performance

Examiners tend to give diagnoses of many of the student interpreters' performances, for good or for bad; this was seen in earlier discussions. The areas of diagnosis range from listening comprehension, to the interpreters' advanced preparation and background knowledge, to the students' training level, to the use of interpreting

strategies and skills, and more. Sometimes examiners even tried to find excuses for the student interpreters, such as saying that they might be tired in the booth, but this was unlikely because the examination task was only the first three minutes of the speech.

Among the five super examiners, for example, PCC2 commented more on the students' use of interpreting skills and strategies, sometimes even making suggestions to the students about how to improve in the future. This could be due to the examiners' professional habit because seven out of the ten PCC2 individual examiners are interpreter teachers (Table 4-9.a). Comment 43 below is a typical example illustrating how the examiners diagnosed students' performance.

> **Comment 43 (translation)[80]:** D missed too much. […] basically I'm not sure if it's because she does not understand. It's really difficult to tell unless she takes a [listening] comprehension test. Overall she lagged behind a lot. […] I feel that this is more like a problem of (lacking) basic practice. […] For example, she probably has never adjusted her pace of speaking. Sometimes under some conditions you can speak slowly, no problem with that, but there are also times when you need to speak faster. I feel that she lacks this kind of judgement (to change pace).

However, diagnosis that does not have the support of evidence, such as a listening comprehension test as mentioned in Comment 43, is unlikely to be reliable because of its subjectivity, which is dependent on the examiners' different experiences, backgrounds, and indeed personal preferences and professional habits.

Therefore, based on the purpose of an interpreting examination, there should perhaps be guidelines on how these kinds of opinions should be considered in the decision-making process, or indeed, how the examiners should be trained and members of an examination panel be selected for good levels of intra- and inter-rater reliability.

---

[80] **Comment 41 in source text Chinese:** D 漏得太多了，[…]基本上，我不太確定他是不是聽不懂啦，這個真的很難說，除非幫他做理解測驗。他整篇就是落後的情況太多，[…] 我覺得這個比較像是基本練習的問題。[…]他從來大概從來沒有調整過自己講話的速度，有時候在某些情況下你可以講得慢一點，沒有問題，有些時候你必須要能夠講的快一點。那我覺得他缺乏這個判斷的能力。

6.5 Summary discussion

This chapter explored what lies behind the consistent and inconsistent judgements among the examiners. This section will summarise and discuss the findings and answer the research question of this chapter, which is the final one stated in 1.3.

Firstly, a range of examiner behaviours have been noted from the observable external behaviour, such as the use of assessment tools, to the internal behaviour, which is less straightforward to observe. As far as the external behaviour is concerned, interpreter and non-interpreter examiners have different approaches to using the assessment tools. When assessing students, interpreter examiners tend to depend on their professional interpreting skills rather than simply to read the speech script; whereas non-interpreter examiners tend to rely more on the speech script to assess students. Some interpreter examiners would even try out the examination tasks to gauge the difficulty of the task before listening to student interpreters' performances. Then they would often assess the student interpreters based on their notes made during the try out and during listening to the student interpreters' performances rather than by using the speech script. Nevertheless, many of both the interpreter and non-interpreter examiners 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 assessment behaviours were inferred from the interview comments of the subject examiners when comparing the student interpreters. The study found that the examiners in general applied similar assessment criteria and 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 Fidelity and Completeness criterion, the examiners would then make the final decision by using the Presentation and Delivery criterion, this process was described earlier 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 over 86% of the decisions were made based on the two primary criteria (5.7).

Although the FCD approach is common to most examiners, variations in judgements do occur. Examiners' assessment strategies vary from one to another. 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. Three main types of internal examiner 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. Examiner behaviour plays an active role in the variations in examiners' judgements.

Simultaneous interpreting imposes complex and high cognitive workloads on both the student interpreters and the examiners; the latter need to observe, assess and make a judgement. Therefore, few examiners are likely to be able to pay equal attention to all the details of every performance and are likely to make the judgement by impression, or by holistic judgement, which is not the most reliable assessment method.

Examiner bias includes the primacy-recency effect and personal preferences for the style of delivering the interpretation. The order in which a number of student interpreters' performances are assessed will affect the way an examiner perceives the interpreting performances. A poor first performance tends to enhance the marks given to later performances. Personal preference for a delivery style 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 subjective viewpoint. All these will play a part in causing inconsistencies in the examiners' judgements.

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 performance assessment. When assessing student interpreters' interpretations, the examiners may refer to their experiences in the field and give different weightings to certain speech content according to the context of interpretation, such as a business meeting. Those interpreters who are also in the teaching profession tend to have different focus 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, including preferred interpreting skills and strategies.

When considering errors and omissions in the interpretation, generally speaking, both interpreter and non-interpreter examiners prefer the use of the condensation norm, or reduction strategy. They would rather that the student interpreters omit a message that is not fully understood, than interpret it incorrectly and cause more confusion. This condensation norm, nonetheless, can only be safely used when an interpreter fully understands the source messages, or is aware of the situation that certain messages are not fully understood, so both major and secondary information can be processed for the reduction strategy.

However, as Sawyer reported in his case study, "the [interpreter] jury members are a heterogeneous group in terms of professional experience as well as experience in teaching and testing" (2004: 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 own professional judgement as discussed above, those differences will play a role in affecting the consistency of their judgements in the interpreting examination. Taking the super examiners in this study as an example, overall they used similar assessment criteria and followed the FCD approach, but their judgement patterns in terms of the student rankings were different.

Therefore, to answer the final research question stated in 1.3, based on the discussions and findings in this chapter, the consistency and inconsistency of the examiners' judgement patterns appear mainly due to the examiners' various assessment behaviours as discussed above.

CHAPTER 7

A Conceptual Model of Interpreting Examinations

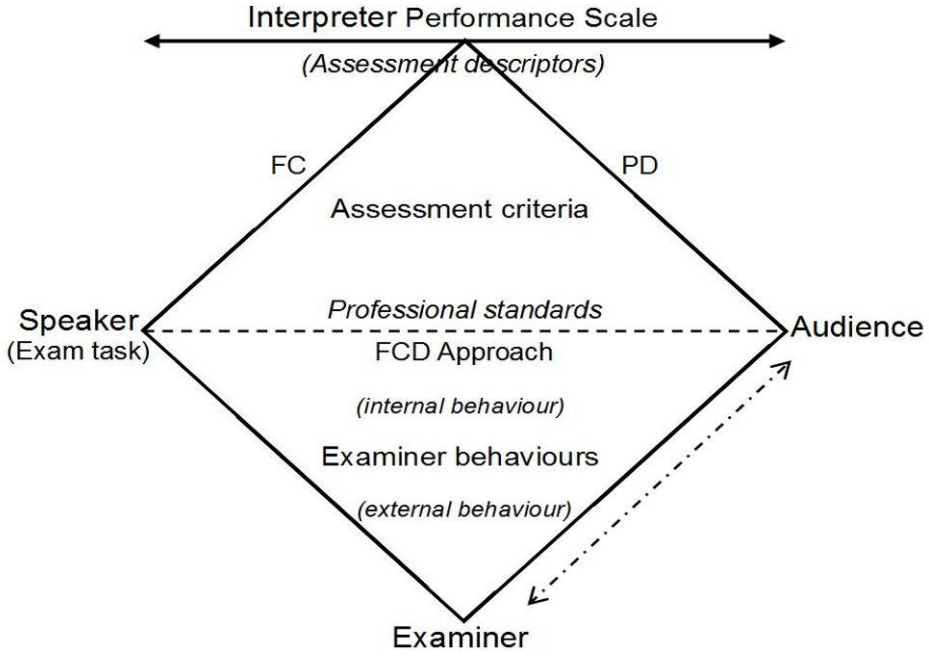
7.0 Preamble

This study explores and investigates how the examiners assess student interpreters in a simulated examination of simultaneous interpreting. As mentioned in Chapters 1 and 2, Sawyer found it "fuzzy" when the interpreter examiners applied their assessment criteria (2004: 185), and variations and inconsistencies were also obvious in interpreter examiners' professional judgements (ibid: 188). Inconsistent judgements (Chapter 4) and, to borrow Sawyer's term, fuzziness in the use of assessment criteria, such as their overlapping and linked conceptual properties (Chapter 5), are also two themes identified in this study. In addition, examiner behaviours identified in this study (Chapter 6) give some clues to an explanation of how the examiners make their judgements, consistent and inconsistent. It is hoped that the findings in this study, therefore, shed some light on such fuzziness and inconsistencies in interpreting examinations.

In this chapter, based on the findings, a conceptual model is proposed to illustrate the relationships between the various elements in a typical interpreting examination and to facilitate further discussions of the findings. In 2.1.2, the main components in the simultaneous interpreting examination of this study were illustrated as the concrete event in Pöchhacker's (2001) multi-perspective figure, which is depicted as a diamond shape (see Figure 2-1). The conceptual model of interpreting examinations (hereafter the IE model) expands and uses this diamond shape to illustrate further the relationships of the finer elements identified in this study. The basic IE model is illustrated in Figure 7-1 and explained below.

In the basic 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),

Figure 7-1 Basic conceptual model of interpreting examinations



FC: Fidelity and Completeness, **PD:** Presentation and Delivery
FCD Approach: Fidelity-Completeness-Delivery Approach

and Presentation and Delivery (PD). The assessment criteria are based on the professional standards (i.e. the base of the assessment criteria triangle) that professional interpreters follow when they interpret between the Speaker (Exam task) and the Audience. Interpreting examinations adopt these standards as assessment criteria for test validity reasons (see 2.1, 2.2 and 2.3).

In Figure 7-1, the factor of examiner behaviours in interpreting examinations is also illustrated by a triangle of Speaker-Audience-Examiner, or the examiner behaviour triangle. At the Speaker-Audience interface adjacent to the assessment criteria triangle is the Fidelity-Completeness-Delivery (FCD) Approach. The Examiner usually follows this general approach to apply the assessment criteria (see 6.3.2), which is influenced by a range of external and internal examiner behaviours.

The IE model is represented as a 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

assessment criteria triangle and the *examiner behaviour* triangle, with the Examiner positioned at the bottom balancing everything on top.

In the *assessment criteria* triangle, or dimension, the Examiner needs to find a balanced weighting of multiple assessment criteria in order to make a judgement (see Chapter 5); whereas in the *examiner 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 (and rating scales if available), while the internal behaviour relates to the Examiner's personal ways of interpreting and receiving the messages based on their own preferences and professional experiences (see Chapter 6). The Examiner also plays a dual role of an assessor as well as a member of the audience (see 5.4), which is indicated by a dotted double-headed arrow in the behaviour dimension. Alternating between the two dimensions, the Examiner attempts to maintain a balance that is intricate and delicate, i.e. to make 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. In other words, 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 or purposes.

With this preamble, sections 7.1 and 7.2 below discuss the two dimensions in the basic IE model above by referring to the theoretical framework of language testing reviewed in Chapter 2, and expand the model as the discussions move on.

7.1 The criteria dimension

This dimension concerns the findings of the five categories of assessment criteria (Chapter 5). As argued in Chapter 2, vague definitions of test construct and assessment criteria open a door to subjective judgements, and therefore, risk leading to inconsistent examination results (see 2.3 and 2.4). In this section, we shall discuss whether the findings of this study can match the claimed test constructs of interpreting and help with better construct definitions (7.1.1); we shall also clarify the fuzzy relationships between the construct components and assessment criteria by referring to a language use model in language testing (7.1.2 and 7.1.3). The objective is to develop and refine the assessment criteria dimension of the IE model for a better design of interpreting assessment, making it easier to measure the test constructs of interpreting.

7.1.1 Matching the constructs and assessment criteria of interpreting

In Chapter 2, some sample construct statements for interpreting examinations were reviewed. Those sample statements share four basic constructs that a good interpreter requires: *language competence*, *interpreting skills*, *background knowledge*, and *personal aptitude*. However, these four constructs lack clear definitions and are regarded as difficult to operationalise (see 2.4.2.a).

A construct can be useful only when it is operational, that is the construct is "associated with 'things' that can be observed, and that these 'things' can be scored" (Fulcher, 2003: 18). In this study, the assessment criteria were extracted from the examiners' verbal comments, the criteria used and articulated by the examiners. In this respect, therefore, when the examiners made comments on how they assessed the

student interpreters, they were, in effect, describing the construct of interpreting. So, to a certain extent, the assessment criteria identified could be considered as operational to measure the test constructs of interpreting examinations.

The five assessment criteria identified in Chapter 5 are summarised and listed with their main conceptual properties in Table 7-1. We shall discuss if the identified assessment criteria can match and make operational the four basic constructs of interpreting as mentioned above. The objective is to test the claims of theory (i.e. the four constructs) with the empirical data in this study, and refine them if necessary.

Table 7-1 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

First, there are two clear matches: *personal aptitude* with the criterion of Foundation Abilities for Interpreting, and *interpreting skills* with the criterion of Interpreting Skills and Strategies. The *background knowledge* construct is also a match, with the resourcefulness conceptual property of Interpreting Skills and Strategies: the

construct of *background knowledge* is usually inferred by observing an interpreter's use of the preparation and anticipation strategies. With the support of background knowledge, the interpreters can do a better job in processing the messages in the source language, arranging them in a way that is more suitable to be delivered in the target language (see 5.5.1). For example, an interpreter can apply the anticipation strategies when he or she is familiar with the subject under discussion in a conference, or already knows the speaker's stance on a topic during a debate. An interpreter can acquire such suitable background knowledge to a certain extent with preparation strategies before a conference interpreting assignment; a better knowledge base enables an interpreter to make better strategic choices that are available onsite (see 2.1.5).

Table 7-2 Four constructs of interpreting and their matching assessment criteria

Constructs	Assessment criteria
personal aptitude:	Foundation Abilities for Interpreting
interpreting skills:	Interpreting Skills and Strategies
background knowledge:	Interpreting Skills and Strategies
language competence:	Presentation and Delivery Audience Point of View Interpreting Skills and Strategies Foundation Abilities for Interpreting

Language competence is an interesting construct as it has clear and direct links with four of the five assessment criteria. It is a perfect match with two criteria – Presentation and Delivery, and Audience Point of View, whose conceptual properties are mainly about the use of language when delivering the messages (see 5.2 and 5.4). This construct can also be found, or observed, in the conceptual properties of the other two criteria: Interpreting Skills and Strategies (paraphrasing, summarising, etc.), and Foundation Abilities for Interpreting (listening comprehension). Such permeation to so many assessment criteria may imply that the construct of *language competence* carries a

lot more weight than those of the other three constructs. It would be over-simplistic, therefore, to view all four constructs as equals in terms of their importance and/or practicality to operationalise in the interpreting examinations. Table 7-2 shows the above matches between the constructs and assessment criteria.

A missing link: the construct of message equivalence

Surprisingly, the four basic constructs appear to have no clear and direct link with the criterion of Fidelity and Completeness (FC), which is missing in Table 7-2. This indicates a gap between the claims of theory and the empirical data in this study. Although discussions about the concept of fidelity may be found in the literatures of interpretation and translation studies, most of them stem from a pedagogical perspective, not for testing purposes. For example, Gile (1995a) dedicated a chapter in his book to discuss the concept of fidelity in translation and interpretation. He prioritised four types of information – Message, Linguistically Induced Information, Personal Information, and Framing Information – to help student translators and interpreters understand that it is the Message that must be reformulated rather than words; the other three types of information are secondary and can be condensed, or omitted if necessary. Even with these information types as guidance, nevertheless, cautions were given that "remedial action" is still needed lest students go "overboard" into adaptation (1995a: 49-74).

In the interpreting studies literature on the test construct of interpreting, Sawyer did mention the FC criterion, saying that the examiners are looking for the examinees' ability to interpret with faithfulness to the meaning and intent of the original (2004: 97), but did not elaborate further (see 2.3.2). In fact, the issue of fidelity and completeness in interpretation studies is left largely unresolved in spite of many product-based research studies, and the main reason is that it is still difficult to objectively quantify and measure the *messages* in the interpretation output (Pöchhacker & Shlesinger, 2002: 251).

As the findings of this study show, and for practical reasons, the assessment of the messages in interpretation needs to rely on subjective professional judgement. The lack of discussion in the existing literatures on regarding FC as the assessment criterion in relation to the test construct of interpreting may be due to the fact that it is usually taken for granted, and that a practical model or methodology is yet to be developed to tackle this issue systematically.

In terms of the interpreting assessment, some examiners may simply consider FC as a criterion to measure the *interpreting skills* construct. However, this study finds that FC conceptual properties are a distinct category from the *interpreting skills* at the operational level in the assessment process, and that FC is the primary assessment criterion that accounts for more than 50% based on which the examiners made their decisions (see 5.7). On the operational level when assessing interpreting, as analysed in 5.3 and 5.5, the FC criterion focuses on three main areas: content accuracy, speaker intention, and contextual consistency by comparing the equivalence of messages between the source and target languages; whereas the criterion of Interpreting Skills and Strategies (ISS) refers to the interpreter students' resourcefulness and multi-tasking ability to utilise various linguistic and cognitive skills to process messages and deliver them in the target language (also see Table 7-1). The ISS criterion is, as found in this study, more like a diagnostic tool during the judgement process for explaining the success or failure of the student interpreters' achieving the FC criterion, i.e. poor ISS leads to unsatisfactory FC, and good FC usually is the result of competent implementation of ISS. The two criteria are closely linked, but they are distinctively different as shown from the analysis of the examiners' verbal reports in the act of comparing and judging student interpreters' performances. The construct of *interpreting skills* as listed in Table 7-2, therefore, does not refer to the equivalence of messages as denoted in the FC criterion, which is missing in the table.

The identified FC assessment criterion in this study, thus, may be pointing to the need to formulate a fifth test construct for interpreting i.e. the ability to interpret with faithfulness as Sawyer put it in one of his sample construct statements. Given the empirical findings in this study and the discussions in the aforementioned literatures, pedagogical and assessment-focused, "message" is the critical component when analysing the ability of interpreting. For the convenience of discussion, therefore, we shall refer this fifth construct of interpreting as *message equivalence*.

In order to use this FC criterion to measure the construct of *message equivalence*, the examiners need to observe and score based on something more tangible than just the spoken messages. In the case of interpreter-mediated communication, languages are vehicles to carry the messages from the speaker to the audience via the interpreter. Therefore, *language competence* becomes a yard stick for applying the FC criterion to measure the construct of *message equivalence*, which looks at three areas of the interpretation: content accuracy, speaker intention, and contextual consistency (Table 7-1). In other words, the examiners need to observe and score interpreting performances by comparing the messages in both the source and the target languages. From this perspective, the construct of *language competence* is directly and indirectly linked to all five of the assessment criteria.

Being the vehicle that carries the messages, *language competence* is fundamental to interpreting. However, some aspects of language competence, such as reading and writing, are not immediately required for interpreting. For instance, reading skills may be useful to interpreters when preparing for a conference interpreting assignment (see 2.1.5), but in the line of work, listening and speaking are more relevant to interpreting. An interpreter first listens to the source speech to receive the messages for interpretation, and the output interpretation is delivered in the form of spoken language, i.e. speaking. Nonetheless, the interpreter's listening comprehension ability, can only be inferred from

observing the output interpretation (see 5.6), which is speaking, the aspect of language competence that may be directly observed and assessed. This is also the reason why the conceptual properties of assessment criteria focus more on speaking (Presentation and Delivery) rather than on listening (Foundation Abilities for Interpreting) because it is easier to observe and infer the competence of *speaking* from the interpreting performances, that is, in relative terms when comparing with *listening*.

The interpreting assessment, nevertheless, is different from assessing language competence per se, as in a listening or speaking test in the discipline of language testing. The added intricacy in the case of the interpreting assessment is that the messages in both the source and target languages need to be compared and assessed from a multitude of perspectives (see 2.1.2 and 5.4.3). The interpreting examination, then, is a combined test of listening and speaking abilities, and more importantly, the ability to transfer messages between two languages, i.e. the construct of *message equivalence*. In the case of simultaneous interpreting, everything has to be done simultaneously, adding a very high level of stress to the work for both the interpreter examinees and the examiners.

Given the analysis above, therefore, it shows that the construct of *language competence* for the interpreting examinations 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 (also see 6.2.2), leading to inconsistencies in the examination results. Therefore, it is necessary to further clarify the intricate relations between the test constructs and the identified assessment criteria.

7.1.2 A theoretical framework for the construct of interpreting

To unravel and clarify the intricate links between the test constructs and assessment criteria in interpreting examinations as discussed above, the analysis strategy used here is first to look at the language-related conceptual properties of the five assessment criteria for interpreting, and then compare them with the constructs of language in the discipline of language testing. Therefore, this section makes use of the knowledge of the language testing discipline in order to clarify the complex relationships between the assessment criteria in interpreting examinations, and to establish what they are measuring and how, i.e. to have better construct definitions.

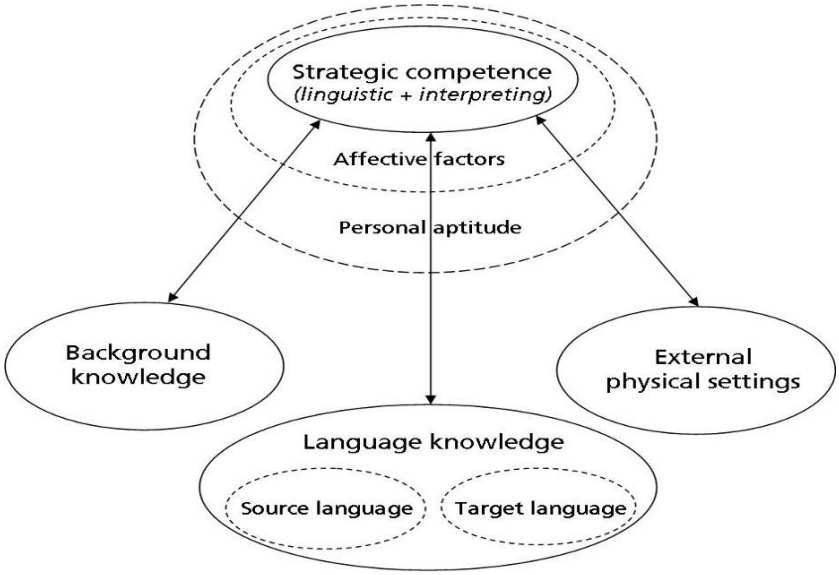
In this study, the construct of *speaking* (i.e. Presentation and Delivery) is mentioned more often than that of *listening* (i.e. Foundation Abilities for Interpreting) in the identified conceptual properties of assessment criteria. This indicates that the construct of *speaking* is easier to operationalise in the interpreting examinations. Therefore, the discussions in the sections below may be conducted more from the perspective of the construct of *speaking*. Nevertheless, *listening* is still an important component of the *language competence* and will also be considered subsequently in 7.1.3.

In Chapter 2, Bachman and Palmer's model of language use (1996: 63) was reviewed. In that model, there is a construct of "strategic competence" that has a cognitive management function (ibid: 70) (see 2.4.2.a and Figure 2-5). With strategic competence, language users are able to set goals in a communicative context and to manage the various functions of language. In order to achieve the goals, such as starting a conversation or requesting information, the language users assess the situation and utilise their "language knowledge" and "topical knowledge" to interact with people in the settings in which the communicative act takes place, e.g. a test task. The model also

takes into consideration the language users' "personal characteristics" and "affective factors".

Interestingly, when applying this language model to analyse the construct components and assessment criteria in the context of interpreting, it is found that they actually map very closely onto each other. Based on Bachman and Palmer's model, a language-based framework for discussing the constructs of interpreting is formulated as shown in Figure 7-2.

Figure 7-2 A framework for discussing the constructs of interpreting



Adapted from (Bachman and Palmer, 1996: 63)

The language-based construct model for interpreting is also communicative and goal-driven. The goal of an interpreter-mediated communication act is to transfer the messages from the speaker to the audience in a reliable manner. To achieve this goal, the interpreter needs to listen to the speaker's source speech (*language knowledge*, Foundation Abilities for Interpreting, i.e. listening), utilise various skills and strategies, including multi-tasking, to process the received messages (*strategic competence*, Interpreting Skills and Strategies), and then deliver the messages to the audience in the target language (*language knowledge*, Presentation and Delivery, i.e. speaking,).

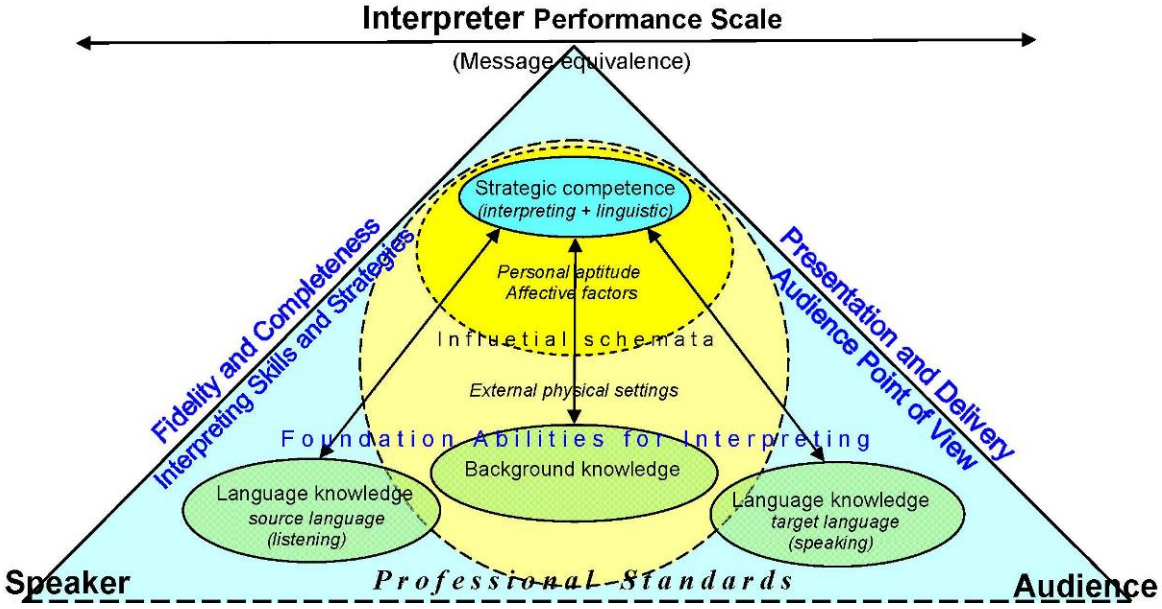
As illustrated in Figure 7-2, those tasks that the interpreters perform are managed by the *strategic competence*, which consists of two components: linguistic and interpreting. The *strategic competence* draws resources from the two knowledge bases – *background knowledge* and *language knowledge* (which includes source and target languages) – in order to complete the multiple tasks simultaneously to fulfil the goal of the communicative act. In the communication process, the interpreters are also influenced by some internal factors, such as the *affective factors* (i.e. personal emotions) and *personal aptitude*. The *external physical settings*, i.e. the environment in which the interpreters work, will also impose some influences on the interpreting performance as reviewed in the professional quality assurance literatures in 2.1. The interactions of these components in the model are indicated by the three double-headed arrows as shown in Figure 7-2.

The overall success or failure of the interpreter-mediated act is then assessed by checking whether the messages are delivered "with faithfulness to the meaning and intent of the original" (Sawyer 2004: 97). In other words, the construct of *message equivalence* is measured by applying the primary criterion of Fidelity and Completeness, which is based on the knowledge of both the source and target languages.

7.1.3 Revise the assessment criteria dimension of the IE model

In the Preamble (7.0), we have illustrated the dimension of assessment criteria as the upper triangle in the basic IE model (Figure 7-1). This section aims to further develop and expand the criteria triangle as part of a revised IE model by using the above language-based construct model for interpreting as a theoretical framework (Figure 7-2), and to further discuss and clarify the interactions of the five assessment criteria in relation to the constructs of interpreting as discussed earlier in 7.1.1.

Figure 7-3 Revised assessment criteria dimension of the IE model



The assessment criteria dimension of the IE model may be revised as in Figure 7-3 to illustrate the intricate relationships between various construct components of interpreting, and how they may be operationalised by means of the assessment criteria identified in this study, which are labelled in blue fonts in the figure.

At the base of the assessment criteria triangle, the three green ovals represent the foundation language competence and background knowledge of an interpreter, i.e. the Foundation Abilities for Interpreting. Their role is like resource reservoirs that support the interpreters to do their jobs. *Background knowledge* relates to the subject areas that interpreters may encounter when working at various conferences; whereas *language knowledge* refers to the foundation competence in the two working languages: source language for receiving, and target language for delivering the messages, i.e. listening and speaking abilities.

The blue oval near the apex of the triangle represents an interpreter's *strategic competence*, which consists of two components: interpreting and linguistic strategic

competences. *Strategic competence* is the theoretical know-how component within the overall interpreting competence. With the goal of interpreting the messages with faithfulness to the meaning and intent of the original, *strategic competence* provides the cognitive management function for an interpreter to simultaneously operate various skills and strategies, of both linguistic and interpreting, to process the messages between the two languages involved. The interactions of the construct components are indicated with the double-headed arrows between the blue and green ovals in Figure 7-3.

The processes and interactions between the blue know-how and the green resource reservoir components are also influenced by some internal and external factors, which are illustrated as two circles in shades of yellow each bordered with broken lines; the inner circle represents the internal influences, and the outer circle indicates the external factor. In Figure 7-2, the factor of *external physical settings*, i.e. the interpreter's work conditions, is a separate component from the internal influences of *personal aptitude* and *affective factors*, i.e. personal characteristics of the interpreter. As they influence the interpreter's performances in one way or another when the *strategic competence* is at work, all three factors, external and internal, are bundled together around the *strategic competence* in the revised assessment criteria dimension as shown in Figure 7-3. This study will refer these factors as the *influential schemata* in the IE model as they provide the basis on which the interpreters work, mentally and physically.

Based on the *professional standards* (see 2.1), the examiners measure the constructs of interpreting by weighting the assessment criteria according to the situational context or topic of the source speech. They will evaluate if the student interpreters' Presentation and Delivery is suitable to the Audience Point of View, diagnose student interpreters' use of Interpreting Skills and Strategies and their Foundation Abilities for Interpreting, and most importantly check whether the interpretation is faithful to the intent of the original by using the Fidelity and

Completeness criterion, i.e. measuring the construct of *message equivalence*. The *message equivalence* construct is indicated on top of the criteria triangle with the Interpreter Performance Scale.

The above assessment activities are mainly operationalised by relying on the two primary criteria: Fidelity and Completeness (FC), Presentation and Delivery (PD). Examiners usually give more weight to the FC criterion than to the PD criterion, i.e. the FCD approach, when making judgements. The examiners also need to be satisfied that any errors and omissions conform to the norm of professional practice, such as the condensation norm, which dictates that secondary and unimportant information can be reduced or omitted under the time constraint of simultaneous interpreting (see 6.4). As for the other three assessment criteria, they are operationalised in a less explicit manner and considered as a whole in the judgement process. These examiners' assessment behaviours will be discussed later in 7.2.

Next, we shall first discuss the key elements of the revised criteria dimension. Figure 7-3 shows the five constructs of interpreting, and the interactions between them in the context of the interpreter-mediated communicative act. The constructs are: *strategic competence*, *influential schemata*, *background knowledge*, *language knowledge*, and *message equivalence*. The sections below will follow this framework to discuss these constructs, referring to Figure 7-3 to illustrate.

7.1.3.a Strategic competence

An interpreter needs to manage and balance various efforts carefully, e.g. listening, processing, and delivering, to achieve successful interpretation. The construct of *strategic competence* is a conceptualised, or hypothetical, know-how component of the overall interpreting competence that enables the interpreters to manage and perform various language-related cognitive tasks at the same time. This study found that the

construct of *strategic competence* is operationlised by the criterion of Interpreting Skills and Strategies, which was presented in Table 7-1 as having two conceptual properties, i.e. resourcefulness and multi-tasking ability, representing both linguistic and interpreting competences (also see 5.5). Therefore, as shown in Figure 7-3, the construct of *strategic competence* has two components: linguistic and interpreting.

The construct of *strategic competence* in the IE model stems from Bachman and Palmer's language use model (1996: 70) (see 2.4.2.a and 7.1.2) which explains the linguistic component in this construct. The linguistic competence here mainly refers to speaking skills in the target language, such as summarising, paraphrasing, self-correction (restructuring), syntactical conversion and alignment, or code switching, between the two working languages (also see 7.1.3.d below).

In the context of interpreting, Gile's Effort Models (1995a: 159-185) may be employed as a useful conceptual framework to discuss the interpreting component in the construct of *strategic competence*. In essence, the Effort Models explain the trade-offs between various efforts required in performing the interpreting tasks. When an effort is saturated and exceeds the overall capacity of an interpreter during simultaneous interpreting, the other efforts will not function properly.

For instance, in this study the examiners noticed a situation when student interpreters sometimes could not hear the messages well, and commented that it was because they might be too busy delivering the previous messages and did not pay enough attention to listen for the in-coming messages, thus leading to unjustified omissions or errors in the interpretation output (see 5.5.1). The interpreting component in the construct of *strategic competence*, therefore, mainly refers to multi-tasking ability, i.e. the ability to manage and perform various linguistic and cognitive tasks that are necessary to receive, process and deliver the messages in real time, such as the aforementioned balanced efforts explained by Gile's Effort Models.

The term "processing" in simultaneous interpreting usually refers to the implementation of interpreting skills and strategies, such as segmenting sentences for easier comprehension, conversion of the messages, opting for more neutral statements or expressions when in doubt, assessing and prioritising the importance of messages for delivery, i.e. applying the condensation norm and reduction strategy (see 5.5 and 6.4.2). When making the multiple efforts to perform these tasks, there is a short time lag between the utterance of the source messages and the interpretation output, i.e. the Ear-Voice Span (EVS). This construct of *strategic competence* is thus usually measured by assessors observing how effectively and efficiently the interpreters manage the EVS as discussed in 5.5.3.

7.1.3.b Influential schemata

As shown in Figure 7-3, *strategic competence* interacts with the other constructs through some internal and external factors, which are referred to as *influential schemata* in the IE model. The internal factors refer to an interpreter's personal experiences, emotionally and professionally, such as the *affective factors* and *personal aptitude*; whereas the external factor refers to the *physical settings* in which an interpreter works.

In this study, however, perhaps due to the brevity of the examination task, few conceptual properties were found to have direct references to the *influential schemata*; only some indirect references were found. They were noticed in some examiners' comments when they were diagnosing the reasons for certain student interpreters' poor or good performances, such as lacking practice or tiredness, or guessing the student interpreters' background in terms of their regional usage of languages (see 6.4.4).

As reviewed in 2.1.3, the professional quality assurance literatures indicate that work conditions for conference interpreters will affect the quality of their performance (AIIC, 2009). However, this study found little references to the *physical settings* apart

from a few comments on the context and environment in which the business speech took place. This may be due to the limitation of this study's focus on interpreting assessment within the educational context and the way the examination simulation was conducted, so the examiners' comments on the schema-related issues were more relevant to the *background knowledge*, which will be discussed later in 7.1.2.c.

Bachman and Palmer also described the *affective factors* as "affective schemata", which "provide the basis on which language users assess, consciously or unconsciously, the characteristics of the language use task and its setting in terms of past emotional experiences in similar contexts" (1996: 65). The *affective factors* in the case of interpreting, therefore, may extend and relate to an interpreter's personal experiences, including interpreting training, as well as his or her personal view on the subject matters under discussion. The influence of the *affective factors* had also been identified by Ivanova (1996) in her study on the nature of expertise in simultaneous interpreting. By using the method of retrospective protocols analysis, Ivanova concluded that the "emotional state"[81] of both expert and novice interpreters on the whole indicated "a very high level of person involvement with the task at hand", which highlighted "the importance of affective factors in SI" (Ivanova, 1996: 45, 47). Thus, in addition to the high level of stress from the cogitative workload in SI, interpreters need to be able to cope with emotionally-involved situation at work, and still to carry out the interpreting job in a professional manner.

The ability to cope with such affective pressures of an interpreting job may also be related to *personal aptitude*. Some studies of the assessment issues in interpreting focused on aptitude tests for admitting potential candidates to interpreter training programmes (Lambert, 1992; Moser-Mercer, 1994; Pöchhacker, 2004; Sawyer, 2004).

[81] The "emotional state" (i.e. the "*internal commentary*" and "*mood*") accounted for 34% of the subcategories in Ivanova's study data (Ivanova, 1996: 45).

Due to the highly stressful nature of the work, *personal aptitude* plays an important role in a conference interpreter's performance. For example, in addition to language and background knowledge, AIIC listed some typical personal traits for those who wish to become a conference interpreter as follows:

- a commitment to helping others communicate
- an interest in and understanding of current affairs, plus an insatiable curiosity
- world experience away from home and school and a broad general education
- the ability to concentrate and focus as a discussion unfolds
- a pleasant speaking voice
- a friendly, collegial attitude
- calm nerves, tact, judgment and a sense of humor
- a willingness to adhere to rules of conduct (e.g. confidentiality)

Source: (AIIC, 2006)

These personal traits or aptitudes give student interpreters a better foundation in terms of their mental preparedness to cope with the abovementioned effective pressures from receiving the interpreting trainings as well as from the work in real life.

Influential schemata vs. test validity for interpreting examinations

When explaining the *response* validity of a test (2.3.2.b), Alderson et at. pointed out that the processes test takers go through and the reasoning they engage in when responding to the test tasks are important indications of what the test is testing (1995: 176). Therefore, in order that a valid test result may be achieved in the case of the interpreting assessment, the *influential schemata* in the IE model (Figure 7-3) need to be considered when designing the interpreting examinations. For example, student interpreters may be asked to take part in a live panel examination, interpreting from a live speaker in front of a live audience, or even being questioned after the interpreting performance, to see to what extent they can cope with the stress or affective pressures.

However, it is difficult to comprehensively measure *personal aptitude* and personal experiences in one interpreting examination, such as those personal traits suggested by

AIIC; they can only be partially inferred from observing the Presentation and Delivery criterion, such as calm or nervous tones in the interpretation. For practical reasons, the time allocated to complete an interpreting test item is usually short compared to the long working hours in the interpreting booths at real-life conferences. So test duration is also a limitation on observing and measuring how the student interpreters can cope with stress and pressure. Some interpreter examiners in this study, for example, commented that they would want to see how the student interpreters performed in a longer speech to see their "gritting power", i.e. the "resilience to stress" as Sawyer indicated in his construct statement (2004: 97). These are indications of the examiners' attempt to know more about the *personal aptitude* of the student interpreters.

So, the question is: when and how should *personal aptitude* be tested or measured? In an ideal situation when students pass a valid aptitude test before being admitted to study interpreting, theoretically speaking it does not need to be measured again during training as the test normally assesses only the trained interpreting abilities. If, however, *personal aptitude* is something that can be nurtured and developed in the training process of becoming a conference interpreter, then more studies on this topic should be conducted to understand how this could be done. This is an aspect of the age-old debate of nature versus nurture, which is outside the scope of this study.

Despite the above limitations, however, the *influential schemata* do affect the way students interpret, or do not interpret. So they need to be taken into consideration when designing examination tasks for assessing interpreting performances; further studies are needed to investigate the *influential schemata* in interpreting examinations.

7.1.3.c Background knowledge

Background knowledge was mentioned in relation to the Interpreting Skills and Strategies criterion in this study (see 5.5). An interpreter relies on background

knowledge relevant to the speech in order to understand and process the messages better. As discussed in 7.1.2 and illustrated in Figure 7-2, topical knowledge provides the information base for all language use. Presuppositions of topical knowledge in test tasks will give an advantage to those test takers who already have the topical knowledge (Bachman and Palmer, 1996: 65), i.e. a knowledge schema. In the context of simultaneous interpreting, therefore, the interpreter may take advantage of applying the anticipation strategy with the support of background knowledge and the right timing when processing messages (see 5.5.1). This may be the reason why the examiners made comments on the *background knowledge* in close relation to the Interpreting Skills and Strategies, i.e. resourcefulness, so that this construct can be made operational in the process of assessment.

Nevertheless, this study found that *background knowledge* plays a more static role in the process of interpreting when compared to *strategic competence*, which plays an active role in the mind of an interpreter to manage various construct components (see 7.1.3.a). To an interpreter, the *background knowledge* is there *not* for enabling the interpreter to initiate new information or messages like a speaker, but for supporting the interpreter's comprehension of the speaker's messages; so in this respect it plays a passive role. As a result, *background knowledge* is separated from Skills and Strategies for Interpreting in the criteria dimension.

As illustrated in Figure 7-3, in the overall competence of interpreting, the role of *background knowledge* may be likened to a resource reservoir that interpreters may call on to obtain support for better performances. The judgement of an interpreter's *background knowledge*, however, can only be inferred by comparing the interpretation output with the source speech, i.e. by using the two primary criteria: Presentation and Delivery for receiving the interpretation, and Fidelity and Completeness for checking sense consistency. The findings in this study are limited in terms of the *background*

knowledge construct, and cannot be over-generalised beyond the three-minute examination task of a single speech topic on business.

For reasons of test validation (content and face validity), it is important to have appropriate examination tasks, i.e. speeches, on appropriate topics to assess student interpreters. The literature reviewed in Chapter 2 indicated that the nature of quality assurance for interpreting is multi-dimensional (2.1), the domain of interpreting is sensitive to various influence from the parties that are involved in the communication act, such as the expectations of the audience, the speaker and the conference organisers (see 2.1.5); it is also sensitive to world knowledge, communication context and the speech topic (see 2.3.2.c). It is commonplace that an interpreter needs to know a little bit of everything to serve as background knowledge. Only with adequate support of knowledge relevant to the subject matter can an interpreter do a good job.

Therefore, in designing examination tasks and developing assessment criteria for interpreting examinations, a subject- or theme-based approach may be considered, such as interpreting for business, for science and technology, or for international organisations (also see 2.3.2.b and 6.5). There are several benefits of doing so. For instance, the *background knowledge* in these specified subject areas may be tested with more details, and the purpose and usefulness of the interpreting examinations may be defined more clearly. In turn, the assessment criteria may also be developed to be more practically useable. These considerations deserve more investigation and studies to improve the reliability and validity of interpreting examinations.

7.1.3.d Language knowledge

Now let us turn our attention to discussing the construct of *language knowledge*, which consists of two components: the source and target languages as shown in Figure 7-3. Interpreters must have a foundation competence and knowledge in both working

languages to do the job. As discussed earlier in 7.1.1, "language competence" links to all of the five assessment criteria, directly and indirectly. This section attempts to clarify further those links to the other construct components (see Figure 7-3) by distinguishing the comprehensive "language competence" in language testing discipline from the *language knowledge* in the IE model, which refers to the foundation linguistic competence and knowledge in the use of both source and target languages.

To assist clarifying the intricate nature of these links, the discussions below will refer to Fulcher's (2003: 48) framework for describing the construct of *speaking* in language testing (also see 2.4.2.a). As shown in Table 2-1, Fulcher's framework consists of five components: language competence, strategic capacity, and three areas of base knowledge in language use: textual, pragmatic, and sociolinguistic (ibid).

Table 2-1 A framework for describing the speaking construct

Language competence	**Strategic capacity**
Phonology • Pronunciation • Stress • Intonation Accuracy • Syntax • Vocabulary • Cohesion Fluency • Hesitation • Repetition • Re-selecting inappropriate words • Re-structuring sentences • Cohesion	Achievement strategies • Overgeneralization • Paraphrase • Word coinage • Restructuring • Cooperative strategies • Code switching • Non-linguistic strategies Avoidance strategies • Formal avoidance • Functional avoidance

Textual knowledge	**Pragmatic knowledge**	**Sociolinguistic knowledge**
The structure of talk • Turn taking • Adjacency pairs • Openings and closings	• Appropriacy • Implicature • Expressing being (e.g. refer to oneself or the others)	• Situational • Topical • Cultural

Source: (Fulcher, 2003: 48)

In Fulcher's framework, language competence includes three aspects: phonology, accuracy, and fluency. One point to clarify first is the definition of the term *accuracy*. The aspect of accuracy in language testing means different things from that in

interpreting assessment. In language testing, accuracy refers to things like the syntax, vocabulary and cohesion of the language use; whereas in the interpreting assessment, the term accuracy normally indicates the sense consistency between the source language and the target language, i.e. the Fidelity and Completeness criterion. In other words, the conceptual properties of *accuracy* between the two disciplines are different, which needs to be noted when defining the language-related construct of interpreting assessment with references in the language testing discipline.

Language knowledge as foundation linguistic competence

In 7.1.3.a, some of the linguistic skills were ascribed to *strategic competence*, which are more relevant to the processing of messages for interpretation, i.e. the know-how component as a part of the overall interpreting competence. The role of those linguistic skills is identical to the "strategic capacity" of language use in Fulcher's framework, which enables the language users to communicate effectively by using strategies like overgeneralization, paraphrase, restructuring, and so on (2003: 48).

In contrast, the construct of *language knowledge* for interpreting here refers more to the foundation competence in the two working languages. For the target language, an interpreter needs to have good foundation competence in, for example, pronunciation, intonation, vocabulary, syntax, idiomatic expressions and cohesive delivery, the description of which is identical to that of "language competence" in Fulcher's framework (ibid). These are the most obvious to be observed and assessed by applying the Presentation and Delivery criterion, which focuses on three aspects: acoustic, word and phrase and flow of information (see 5.2). As for the source language competence, i.e. listening comprehension ability, this study found that its assessment was operationalised by diagnosing and inferring from the comparison between the output interpretations and the source speech messages (*cf*. 7.1.3.c).

Therefore, speaking overall in the case of interpreting assessment, to operationalise and measure the construct of *language knowledge*, the examiners need to use a combination of two criteria: Presentation and Delivery, and Fidelity and Completeness when assessing the student interpreters' interpretation output. If the Presentation and Delivery is of good quality, i.e. the speaking of the target language, but the Fidelity and Completeness is poor, the judgement, or diagnosis, may be that the student interpreter's listening comprehension of the source speech was inadequate (5.6.1).

Problems of listening comprehension in simultaneous interpreting, however, may result from a number of different causes, such as a low *strategic competence*, insufficient *background knowledge*, or simply because of inadequate *language knowledge*, or a combination difficulty with those construct components. So, when diagnosing student interpreters' performances, examiners sometimes also make a distinction between not hearing the sentences or messages and not understanding the sentences or messages (see 6.4.3 and 6.4.4). In the case of not hearing, for example, it may be that the student interpreter was too busy in processing and delivering the messages, and had insufficient capacity to listen to the source speech, i.e. poor multi-tasking ability *(strategic competence*) (also see the Effort Models in Gile, 1995a). In the case of not understanding, it may be that the student interpreter's source language competence, i.e. listening, and background knowledge is inadequate, so the sentences of the source speech may be heard but not understood by the interpreter *(language knowledge* and *background knowledge*).

Language knowledge vs. background knowledge

This brings us to discuss the differences and similarities between *language knowledge* and *background knowledge*. As discussed in 7.1.3.c, *background knowledge* helps with the use of the anticipation strategy, and is more concerned with subject area

knowledge, such as science and technology, or finance and business. In contrast, the construct of *language knowledge* here refers to the pragmatic and sociolinguistic knowledge of language use within a specific context, such as a keynote speech or a domain specific speech act. In other words, the construct of *language knowledge* in the IE model includes the aforementioned foundation linguistic competence and the base knowledge in language use, such as textual, pragmatic, and sociolinguistic in Fulcher's framework (2003: 48).

The pragmatic and sociolinguistic parts of *language knowledge* in some way also play a similar role of knowledge schema as the construct of *background knowledge* in supporting the interpreters' *strategic competence*. As mentioned before, the goal of an interpreting act is to convey the speaker's messages with faithfulness to the meaning and intent of the original. It has been noted that the speaker's intention may usually be identified in a direct or indirect speech act, especially in conferences and meetings where the language use is formal. In speech acts, language use abounds with fixed expressions that have domain specific functions (Levin, 2003). For example, expressions like "Ladies and gentleman", "It is a great honour for me to…", etc. are commonly heard in the opening of a speech; in closing remarks of a speech, the expression like "On behalf of all delegates" often indicates the speaker's intent to thank the conference organisers for their efforts. Knowing the structure of talk, language use and familiarity with such fixed expressions, will help an interpreter utilise the anticipation strategy. Such familiarity allows the interpreter to reserve more capacity for the strategic processing and delivery efforts to convey the messages, enhancing the overall interpreting performance.

An analogy of the role of fixed expressions in speech acts would be like the function of road signs, which help car drivers to anticipate the route or road conditions ahead and prepare to react. In this sense, therefore, the role of *language knowledge* is

similar to *background knowledge*, i.e. a resource reservoir of knowledge schema for supporting the interpreters' *strategic competence* (*cf.* 7.1.3.c). This strategic support is especially useful in simultaneous interpreting when the speed of processing the messages is high, like a car driving on a high-speed motorway. The examination task in this study, for instance, is to interpret in real time the opening part of a keynote speech at a business conference (3.2.2). To do well in this examination task, student interpreters need to have the *background knowledge* in business and industry, and the *language knowledge* relevant to the speech act, in this case opening remarks and the way a senior business manager gives a presentation at a conference, i.e. the pragmatic and sociolinguistic knowledge in language use.

In summary, so far we have ascribed the "strategic capacity" part of language use in Fulcher's framework and Gile's Effort Models to the *strategic competence* in the IE model as the know-how component that enables an interpreter to manage and balance multiple efforts in performing various language-related cognitive tasks (7.1.3.a); we also discussed the construct of *language knowledge*, including the foundation linguistic competence, and its role as knowledge schema like the *background knowledge* to support an interpreter's *strategic competence* (7.1.3.c and 7.1.3.d). The *influential schemata* were also discussed in relation to the test validity of the interpreting examinations (7.1.3.b). All these constructs of interpreting, as discussed above, are inter-linked as illustrated in criteria triangle of the IE model (Figure 7-3). Their common goal is to enable the interpreters to successfully convey the source messages from the speaker to the target audience, i.e. the construct of *message equivalence*.

7.1.3.e Message equivalence and the interpreter performance scale

As shown in Figure 7-3, the construct of *message equivalence* is indicated at the apex of the criteria triangle with the Interpreter Performance Scale. In 7.1.1 and 7.1.2

we have also explicated that the construct is operationalised by the criterion of Fidelity and Completeness, which compares the messages in the source language with those in the target language interpretation. To successfully achieve *message equivalence* in interpretation, efforts need to be made to carefully balance all the components in the criteria dimension of the IE model.

At operational level, this study found that two primary criteria are used to measure the overall construct of interpreting (5.7), which are Fidelity and Completeness (FC), and Presentation and Delivery (PD) as indicated on the two slops of the criteria triangle (Figure 7-3). Based on the context of the speech and its subject area, the examiners assess the interpretation to see if the messages are conveyed successfully. Different weightings between the two criteria may be considered, depending on the nature and purpose of the communication act. For example, if the intent is to transmit factual information, the FC criterion dominates; if the speech is motivational, the PD criterion may be weighted more because the intent of the *messages* requires high flexibility in the use of the target language to achieve the equivalent effect as in the source language (*cf.* 7.1.3.c regarding theme-based design of interpreting examinations).

The open-ended Interpreter Performance Scale accommodates this dynamic consideration in measuring the construct of *message equivalence*. A "better" interpreting performance may be positioned on any position toward either ends of the scale; for example, for factual and informative type of speeches, the closer toward the Speaker the better, and for motivational type of speeches, toward the Audience. The factual information in an informative speech would be blurred by an interpreter's exciting delivery tones in the target language; whereas the true messages, i.e. the intent to motivate the audience, say in a product promotion conference, would likely be obstructed by the interpreter's monotonous delivery tone even when every word and phrase of the source speech is accurately translated.

The quality of interpreting, therefore, needs to be assessed from a perspective that is not only multi-dimensional (2.1), but also dynamic as discussed here. The interpreting examination tasks and assessment descriptors that accompany the Interpreter Performance Scale need to be developed by taking this perspective into consideration.

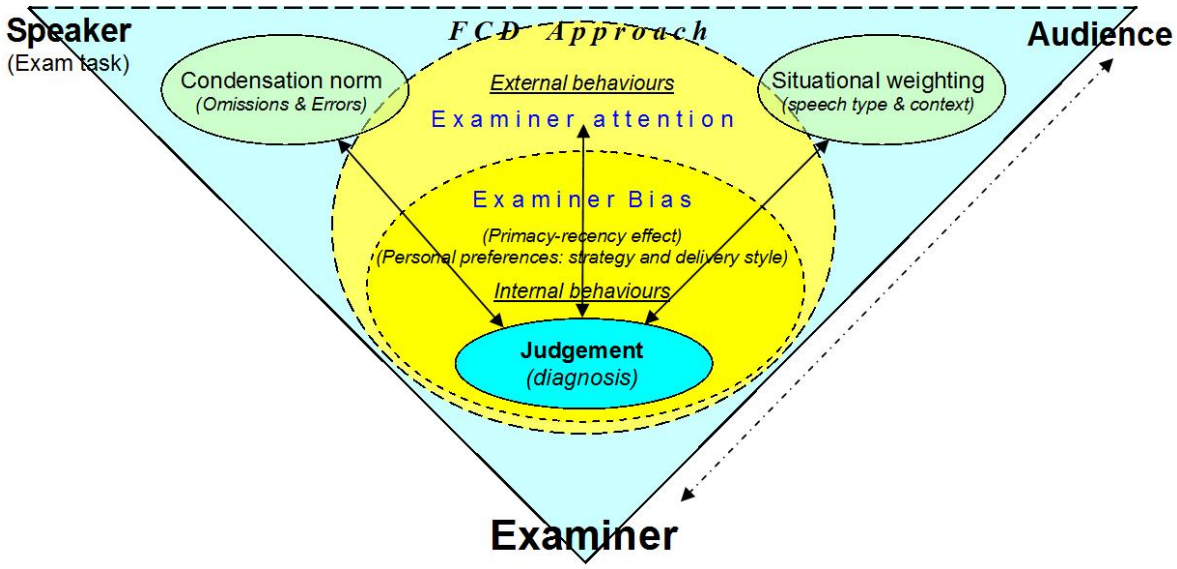
* * *

For formative assessments, the discussions in section 7.1 about the test constructs and assessment criteria, as well as how they interact with each other, are useful guidance to give feedbacks to students for identifying areas to improve. However, in actual operation, there is still little consensus on how some of the construct components could be measured, or if they should be tested, especially for those constructs that can only be assessed by inference, such as listening comprehension, background knowledge, as well as the affective factors and personal aptitudes in the *influential schemata*. More objective parameters are needed to make the assessment more practical and consistent to operate, such as developing rating scales with clear assessment criteria descriptors for the interpreting examinations. These considerations are more relevant to the examiners' assessment behaviours, which will be discussed next in 7.2.

7.2 The behaviour dimension

In Chapter 6, we found that the consistency and inconsistency of the examiners' judgements are mainly due to the examiners' various assessment behaviours. In the Preamble (7.0), the dimension of examiner assessment behaviour has been illustrated as the lower triangle in the IE model (Figure 7-1), which provides a basic framework for discussion. This section discusses the examiners' assessment behaviours with an objective to further develop and expand the behaviour triangle of the IE model.

Figure 7-4 Revised behaviour dimension of the IE model



Based on the findings in Chapter 6, the behaviour triangle can be revised as shown in Figure 7-4, which illustrates in more details how the assessment behaviours identified relate to and interact with each other in the conceptual model of interpreting examinations. The judgement (or diagnosis for formative assessments) of the interpreting performance is influenced by various types of behaviours in the process of assessment, including a general *Fidelity-Completeness-Delivery (FCD)* judgement

approach (see 6.1), and two professionally-referenced behaviours – the *condensation norm* (i.e. interpreter's reduction strategy), and *situational weighting* of the source speech type and context (see 6.4). These assessment behaviours have a direct impact on the use of assessment criteria when making judgements. Other factors that may also affect the judgements are *examiner attention* (see 6.2) and *examiner bias*, which includes the primacy-recency effect and personal preferences (see 6.3). In addition, the examiners may assess the students from the audience point of view, thus, playing a dual role in an interpreting examination (see 5.4).

As noted before, the assessment behaviours range between the observable external behaviour and the less straightforward internal behaviour. The discussions below will be based on these two broad types of behaviours, referring to Figure 7-4 for illustration.

7.2.1 Internal assessment behaviour

The internal assessment behaviour relates to the examiners' ways of interpreting as well as receiving and perceiving the messages based on their professional experiences, and on their personal preferences as individuals.

7.2.1.a FCD approach and professionally-referenced behaviours

The general *FCD approach* and professionally referenced behaviours, i.e. *condensation norm* and *situational weighting*, are illustrated at the upper part of the behaviour triangle in Figure 7-4, which is close to the *Professional standard* in the criteria dimension (Figure 7-3). They may be considered as the examiners' collective assessment behaviours in the interpreting examinations.

This study has identified two primary assessment criteria that the examiners used when assessing student interpreters (5.7), i.e. Fidelity and Completeness (FC), and

Presentation and Delivery (PD). The examiners generally follow the *FCD approach* when assessing student interpreters, i.e. the FC criterion will first be considered, and when it cannot help the examiners to make a satisfactory judgement, the PD criterion will be used (6.1.2 and 6.5). FC is also weighted more than PD when both criteria are considered for making a judgement.

In the judgement process, the examiners would also consider the speech type and context, and apply different weightings to the assessment criteria accordingly, i.e. the *situational weighting* as shown in Figure 7-4. In this study, for example, the source speech is about business so numbers and business terminology are weighted more than the other messages when assessing the student interpreters' performances.

In addition, the examiners would also follow the *condensation norm* to consider the weightings of omissions and errors when applying the assessment criteria of Fidelity and Completeness. For example, due to time constraints in simultaneous interpreting, it is acceptable that the secondary and less important information in the source messages may be skipped or reduced in the interpreters' output interpretation, i.e. the reduction strategy (see also 6.4.1, 6.4.2, and *cf.* 7.1.3.a).

According to the examiners' comments (Chapter 6), the above assessment behaviours are shaped and formed by the examiners' professional experience of interpreting. The results of the interpreting examinations are thus based on these professional judgements, which is an important element of test validity (see 2.3.2.b).

When most examiners follow a similar approach and a shared professional norm to assess the student interpreters' performances, the overall between-examiner consistency of the examination results may be maintained. However, as Sawyer pointed out, the background of the interpreter examiners varies and inconsistencies in their judgements are bound to happen, which is a cause for concern (2004: 184). The findings of this study also show such inconsistencies; the five super examiners (made up of thirty

examiners) show five different judgement patterns in terms of student rankings (see 4.3), and variations in judgements are also obvious between individual examiners.

However, this study could not find a clear relationship between the examiners' judgement patterns and their backgrounds. Although different judgement patterns were found between the two large clusters of examiners (i.e. PCC1 and PCC2 super examiners) that are consisted of largely non-interpreter and interpreter examiners respectively, there are still inconsistencies between the interpreter examiners themselves in the other clusters (see 4.4).

For example, market-oriented interpreter examiners tend to weight the Presentation and Delivery more than the examiners teaching in the universities do; interpreter teachers tend to consider more assessment criteria and try to give a diagnosis of student interpreters' performances (see 6.4.1 and 6.4.2). At the same time, some non-interpreter examiners also share similar judgement patterns to the interpreter examiners in all the five clusters (see 4.4 and 4.5).

Therefore, it appears that more factors than just the examiners' professional background will cause inconsistencies in their judgements. This study has identified two types of such factors – *examiner bias* and *examiner attention*, which are illustrated as the inner two circles with broken lines in Figure 7-4. The way these factors influence an individual examiner's judgement is similar to the way the *influential schemata* affect an interpreter's performance in the criteria dimension (see 7.1.3.b). In other words, these two types of assessment behaviours are more personally related to the examiners, which will be discussed in the sections below.

7.2.1.b Examiner bias

As shown in Figure 7-4, the inner circle of *examiner bias* includes two biases identified: the primacy-recency effect and personal preferences (see 6.3). The examiners

will be influenced by these biases, consciously or unconsciously, when judging or diagnosing student interpreters' performances.

As noted in 6.3.1, the primacy-recency effect is a cognitive bias and may create a structural problem for interpreting examinations. The order of student interpreters being assessed will affect the way an examiner perceive their performances, especially when a poor performance is followed by a better one, or vice versa. This examiner behaviour in performance assessment has been researched and known in other disciplines such as psychology (see Steiner & Rain, 1989). In the case of the performance-based interpreting assessment, this cognitive effect still needs further study to determine to what extent it affects the examiners' judgement, particularly in a panel examination where many examinees are being assessed.

The other identified *examiner bias* is the examiners' preferences. This study identified two types of preferences – (1) the preference of interpretation delivery style, and (2) the preference of the way the interpretation is done, i.e. the examiners may have their own preferred interpreting strategies. The delivery style preference is mainly perceived from the audience point of view, whereas the preference of interpreting strategies is concerned more with an examiner's own professional habits of interpreting (*cf.* 7.2.1.a). For example, how a sentence is segmented when being simultaneously interpreted into another language with a different grammatical sentence structure, and the management of Ear-Voice Span (i.e. how far to lag behind the speaker) when processing messages with different level of complexity or delivery speed of the speech. These preferences will affect the examiners' judgements when they assess the student interpreters from a dual role perspective as Examiner and Audience as indicated in Figure 7-4 (also see 5.4.3, 6.3.2, 6.4.3).

In this study, many examiners were unaware of being influenced by the viewing order of the student interpreters until they reviewed the recordings; some examiners

changed their minds or adjusted their comments after the second or third reviews of the examination recordings. A few examiners had strong personal preferences for the delivery style and strategies in interpretation; they were aware of their preferences and made their decisions accordingly. In one way or the other, these assessment behaviours contributed to the inconsistent examination results found in this study. In some cases, a few examiners even made decisions that contradicted themselves during the paired comparisons, i.e. intra-rater inconsistency (see 4.4.3).

To reduce the influence of the *examiner bias* such as mentioned above, we may learn some useful experiences from the field of language testing. In language testing, the training of examiners, or rater training, is used to ameliorate the problem of random error in the examiners' judgement (see reviews in 2.5.1). However, examiner training can only reduce "extreme differences" in assessment behaviours and the examiner variability cannot be totally eliminated (Lumley & McNamara, 1993: 3). Researchers in language testing, therefore, hold the view that the function of the training of examiners is to train raters to be more self-consistent, allowing for some variability in rater reactions to the test performances (Weigle, 1998: 265), i.e. the examiners can have some room to assess in a natural way based on their professional judgement. In order to do so, sub-patterns in the behaviour of examiners need to be identified for compensation in the test design (Lumley & McNamara, 1993: 3).

In the case of the interpreting examinations, therefore, the findings of this study are useful pointers to the design of examiner trainings for improving the examiners' self-consistent level of their judgements, and to the development of better examination procedures that help avoid or minimise the potential harm from the *examiner bias*.

7.2.1.c Examiner attention

As mentioned in 1.1.2, the complexity of the SI task imposes high cognitive demands on interpreters and examiners alike. When assessing simultaneous interpreting, just as an interpreter must, an examiner needs to multi-task, paying attention to a number of assessment details at the same time. Examiners need to listen to the interpretation, compare the messages with the source speech, make notes of any errors and overly literal interpreting of the source speech, and make a judgement of the interpreting proficiency by taking into account the various assessment criteria. All these tasks impose a high level of stress on the examiner's concentration and memory load.

When there are many student interpreters to be assessed, examiners may not be able to note and remember every detail of every student interpreter's performance (see 6.2), especially when in a live panel examination. That is why many examiners take notes or review the examination recordings to help make better judgements. Even so, many examiners in this study needed to review the examination recordings (some up to three times), or to consult the speech script again before making a decision. In some cases a decision was reversed after reviewing the scripts and recordings. The need to review recordings and notes indicates that there is a limit to an examiner's attention span and memory load in a simultaneous interpreting examination.

Given the complexity of the criteria dimension discussed in 7.1, the examiners may often resort to holistic marking as a result (see 6.2.2) or pay more attention to one criterion or less to another, depending on their attention span as well as personal preference and bias as discussed in 7.2.1.b. All these factors combined together make it difficult to maintain a good consistency level of judgements between or even within individual examiners. At an examination panel when there are divergent opinions, therefore, it is important that the jury discussions are evidence-based. Deliberations that

are based only on subjective judgements with no evidential support may often lead to certain unconvincing results among the juries. The jury discussions may be further complicated when there are examiners "who remit to the learning process and results obtained during the year (instead of evaluating the performance during the exam), who want to impose their own personal view, or who think they wield more prestige and thus should have a decisive vote" (Vermeiren, 2010: 297).

Clearly, the outcome of jury discussions may be intervened by some factors, such as the holistic and subjective judgement of examiners, who unavoidably have certain *examiner bias* as discussed previously. Under such circumstance when holistic and subjective judgement is inevitable, one way to facilitate this judgement approach is making use of appropriate assessment tools and procedures to compensate for the limitations in the examiners' attention span and memory load, such as using speech script to assist the examiners' note-taking while listening to the student interpreters' performances. Then, the examiners' notes can be regarded as a form of assessment evidence for jury discussions (Liu et al., 2008: 19). With an evidence-based discussion, it may reduce the level of unnecessary interventions from *examiner bias*. These considerations are related to the external assessment behaviour of examiners.

7.2.2 External assessment behaviour

The external behaviour mainly concerns the use of assessment tools. As discussed above, working under high cognitive and memory load leads to the examiners having uneven attention spans, with the likely result being inconsistent judgements. The solution to this problem, therefore, is to give the examiners support in making assessments through using practical assessment instruments like the source speech script for note taking and a rating scale with clearly defined assessment criteria.

Regardless of the examiners' background, interpreter or non-interpreter, this study found that using a speech script for note taking generally helped raise the consistency level of the examiners' judgements (see 4.2.2). Using a source speech script for note taking helps the examiners by reducing the cognitive and memory load involved in listening to both the source and target speech at the same time, and eases the checking of the primary criterion of Fidelity and Completeness objectively.

Despite the benefit of using a speech script, not every examiner in this study used one, and among those who did use the script for note taking and assessment, there was some variation in approach. Some examiners just read the script as they listened, while the others took notes with varying degrees of detail (see Appendix B). If the examiners' notes are to be treated as evidence for jury discussion, certain guidelines need to be developed for examiner trainings to reduce the variations in using the assessment tools.

Some examiners also rehearsed the interpreting task before assessing students' performances, which is not uncommon in professional interpreting examinations (Yang, 2000: 162). The main purpose of doing so is to make sure that the difficulty level of the task is appropriate, and that the examiners are aware of where the difficulties of the task lie. Although the rehearsal remains subjective in nature, it allows the examiners to think and comment on the usefulness and validity of the examination task for the benefit of assessment (Vermeiren, 2010: 295). So the rehearsal practice should still be encouraged when setting the examination tasks.

However, according to both Yang's (2000: 162) and Vermerien's (2010: 295) descriptions of the administration of interpreting examinations, and my own experience of serving as a member of panel examiners, the rehearsal practice and the discussion of the suitability of the examination task might only happen shortly before the interpreting examinations, on the same day or a day before the examinations. This leaves very little time, if any, to improve or change the examination task if the difficulty level of the

examination task is found to be inadequate. Thus, when forced to use a less-than-ideal examination task, the examiners often have to adjust the severity or leniency of their judgement when assessing the interpreting performances.

The main benefit of using the practice of last-minute rehearsal of the examination task, such as the above, is its practicality. The between-examiner reliability of the specific interpreting examination may still be maintained, that is, assuming all examiners in the jury panel join the rehearsal. Nevertheless, as a result, this practice of last-minute rehearsal would make it hard to maintain the difficulty level of test items between examinations (i.e. internal consistency of test), and the generalisation of the examination results over time (test stability) would be difficult to ascertain. Adding the risk factor of examiner's reliability, all three criteria (i.e. examiner, internal consistency, and test stability) to evaluate a test's overall reliability are threatened (also see 2.3.3).

In order to alleviate the threat to the test reliability, therefore, the more appropriate timing to carry out the rehearsal practice should be during the test design stage well in advance of the actual examinations. By doing so, it leaves more time to improve the examination tasks when necessary. In the mean time, a consensus among the examiners on the use of assessment criteria also needs to be built to minimise inconsistency. Even with the assessment tools mentioned above, some standardised approach to using them, through examiner training, is required in order to achieve more consistent and reliable judgement results. Understanding the examiners' behaviours discussed in the behaviour dimension in this study, is useful knowledge when designing and formulating examiner training sessions for interpreting examinations.

* * *

This chapter has discussed various elements and components in the workings of interpreting examinations, and proposed a conceptual IE model to illustrate them. Using the IE model as a framework for discussion, a better picture of the intricate relationships between the elements and components emerged, and the two dimensions of the IE model were revised accordingly. Through the discussions, we have gained a better view of the test constructs and assessment criteria for interpreting examinations, and how they may or may not be operationalised. We also have a better understanding of the examiners' assessment behaviours in interpreting examinations. With this knowledge, better interpreting examinations could be developed and improved.

CHAPTER 8

Toward a Better Interpreting Assessment

8.0 Overview

This research study is an exploratory endeavour to understand how examiners perceive the interpreting performances and make judgments in a simultaneous interpreting examination. The results of this study are mostly descriptive in nature, but they yield useful insights to better understand how we assess student interpreters. The findings of this study have been presented and analysed in detail in the previous four chapters – Chapters 4 to 7.

This final chapter firstly summarises the exploration process and findings of the study by reviewing and answering the research questions set out in Chapter 1, and by presenting the complete revised conceptual model of the interpreting examinations (8.1). Following a discussion on the limitations of the study (8.2), suggestions for developing the interpreting examinations are made by presenting a working document of construct specifications for the interpreting examinations (8.3). Some suggestions for the future studies are also made toward the end of the chapter (8.4) with a final remark to conclude this research journey (8.5).

8.1 Summary of study findings

This study aimed to understand how examiners perceive interpreting performances and make judgments in a simultaneous interpreting examination. In achieving the overall aim, the first objective of this study was to identify and determine if there are different judgement patterns of the examiners when assessing simultaneous interpreting. The study found that the thirty participant examiners' consistency level of judgement as a group could be considered as excellent. However, obvious between-examiner inconsistencies were also observed, which raises concerns in terms of the reliability of the interpreting examinations (4.1).

The second research question was to discover whether interpreter examiners and non-interpreter examiners could achieve similarly consistent judgements in a simultaneous interpreting examination. Based on the judgement results of the paired comparisons, it was found that the non-interpreter examiners achieved better consistency level in their judgements than the interpreter examiners did in the simulated examination of simultaneous interpreting in this study. Nevertheless, the reason to the difference in the judgement consistency levels could not be clearly attributed to the pre-determined categories of the examiners' professional backgrounds (4.2). Therefore, using the method of cluster analysis, this study further extrapolated five different judgement patterns that are internally consistent (4.3) to analyse the qualitative data to answer the third research question of this study, i.e. to find out what the examiners' most important assessment criteria are when judging students' interpreting performances.

The study identified six distinct categories of concepts that the examiners commented during their judgement processes. Five of those conceptual categories could be regarded as the assessment criteria for the interpreting examinations, which are

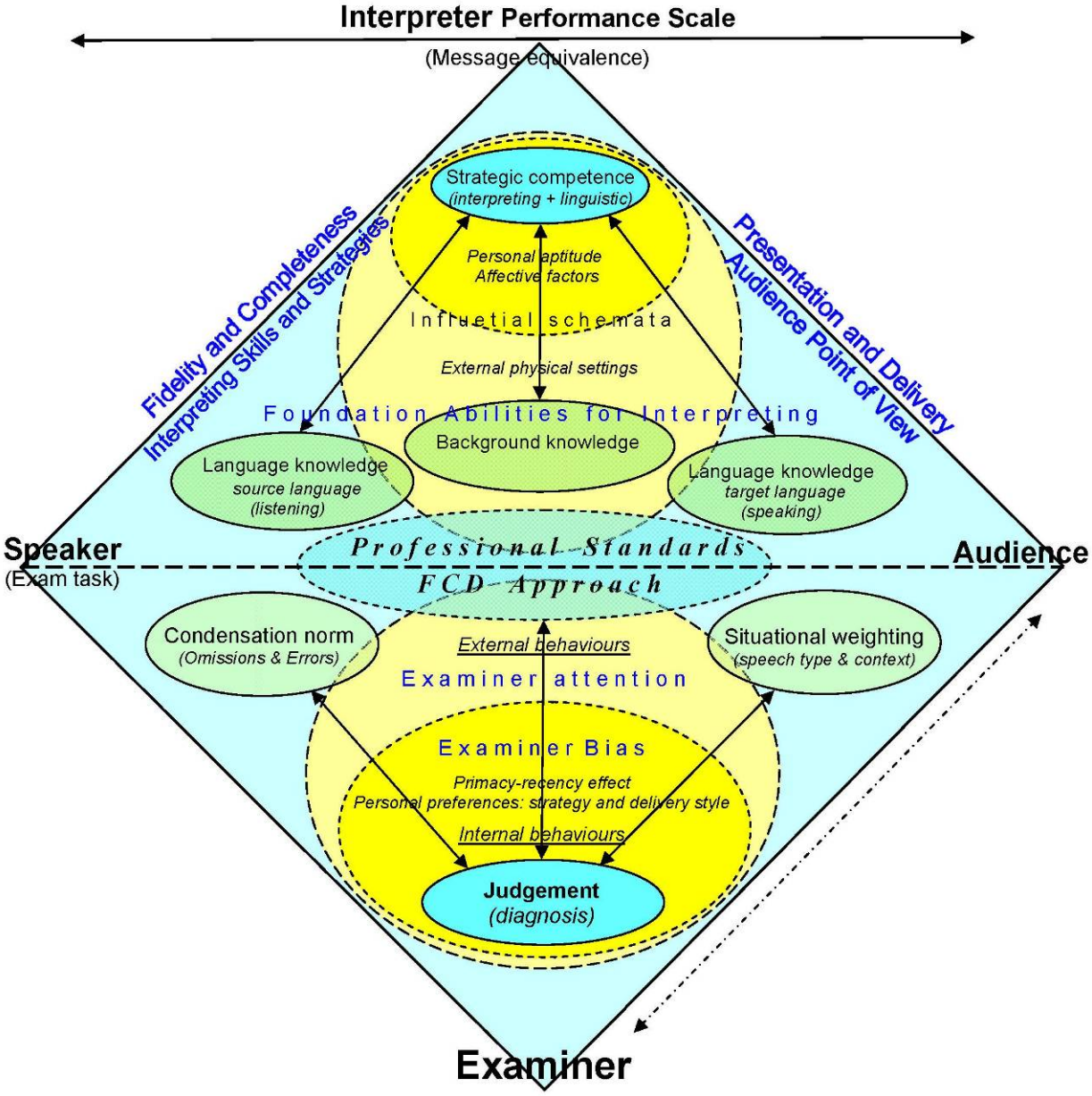
Presentation and Delivery, Fidelity and Completeness, Audience Point of View, Interpreting Skills and Strategies, and Foundation Abilities for Interpreting (Chapter 5). Since these criteria were derived from the comments and descriptions of the participant examiners' when they were judging the student interpreters' performances, they could be considered as criteria that are operationalised to measure the construct of simultaneous interpreting.

From these assessment criteria, therefore, some underlying constructs in simultaneous interpreting assessment were derived or conceptualised, such as *strategic competence* (linguistic and interpreting), *language knowledge* (source and target languages), *background knowledge*, and *message equivalence*; some *influential schemata* were also noticed that may mentally and physically influence the interpreters' performances, such as personal aptitudes, affective factors, and the external physical settings of the examination task (see 7.1).

As for the sixth conceptual category, it was found to be related to a range of factors that impose influences on the examiners' judgements, i.e. examiners' assessment behaviours (Chapter 6). The assessment behaviours could be broadly indicated as two types: external behaviour, such as the use of assessment tools, and internal behaviour, such as *examiner bias* and *examiner attention*. It was also found that examiners would usually take the *FCD approach* when applying the assessment criteria to judge student interpreters based on the *Professional standard*.

This study has confirmed with empirical results that variations in interpreter examiners' professional judgements are evident (see 4.1), which was also noted by Sawyer who observed that the interpreter examiners "make numerous references to various criteria for assessment," which were "fuzzy" in the assessment process (2004: 185-188). The remaining objectives of this study aimed to clarify this fuzziness, or if it could not be clarified, to understand its nature, i.e. to ascertain the relations between the

Figure 8-1 The conceptual model of interpreting examination



examiners' application of assessment criteria and the results of their judgement (the 4th research objective), and to understand the examiners' assessment behaviours that lie behind their judgements, consistent as well as inconsistent (the 5th research objective). In achieving these objectives, this study has formulated a conceptual model of the interpreting examinations, or the IE model, as shown in Figure 8-1.

The multidimensional and dynamic perspective, and the multiple numbers of variables that have to be kept in check by the test designers and examiners, are the

causes of the fuzziness in assessing the quality of interpreting. In addition, the intertwined relations between the test constructs and the assessment criteria for the interpreting examinations make it even more difficult to see clearly how the interpreting assessment can be conducted in a reliable and valid manner. At the centre of this fuzziness is the language-related variable, or language competence, which permeates all of the five identified assessment criteria and their matching test constructs, directly and indirectly. These intertwined relations also lead to a higher level of inconsistency in the judgement of interpreting examinations.

As discussed and explicated in Chapter 7, the IE model is formulated based on theoretical frameworks across the disciplines of language testing and interpreting assessment (Figures 2-1, 2-5, 7-2). It provides a conceptual map to guide us through the fuzziness of the intricate relations between various elements in the interpreting examinations, including the test constructs, the assessment criteria, and the assessment behaviours of examiners. The criteria dimension illustrates how the test constructs and assessment criteria interact with one another; the behaviour dimension depicts how the examiners judge and/or diagnosis student interpreters' performances through some external and internal influences. Balancing all these, the examiners then place the student interpreters on the Interpreter Performance Scale by weighing various criteria according to the nature of the speech test item (Figure 8-1).

Finally, as indicated in the two dimensions of the IE model, the themes or topical subjects of the speech task for interpreting may affect the weightings of the assessment criteria. Therefore, the open-end design of the Interpreter Performance Scale allows the test developer flexibility to design performance scales according to various speech contexts and subject themes with appropriate weightings of the assessment criteria so that a balanced interpreting examination that is both reliable and valid can be achieved.

8.2 Limitations of the study

As with all research, there are limitations to this study in terms of study design, and in terms of the extent to which the findings can be generalised to a broader context of the assessment of interpreting performances.

In real life, interpreting examination panels usually consist of multiple examiners. Under the wider real-life context, therefore, one obvious limitation in terms of the research design is that this study only examined the examiners' judgement process individually, and did not produce data on the interactions and influences among examiners that may be present like in an 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.

Nonetheless, 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. Many interpreter teachers teaching in universities are often required to assess their students on their own from examination recordings (Liu et al., 2008:31). The findings of this study also give useful pointers in understanding how individual teachers may assess student interpreters, and how an improved examination and marking procedure may be developed to help achieve a more reliable result of the interpreting examination.

Another limitation is the examination task. As discussed in 7.1, the findings in this study are limited in terms of some identified construct components, such as the *influential schemata* (e.g. personal aptitudes and affective factors) and the *background knowledge* – i.e. those most closely linked to the test piece, a three-minute examination task of a single speech topic on business. The relatively small numbers of student interpreters and examiners in this study have also imposed limitations on the scope of examiner behaviours identified, such as the rater effects in language testing where larger number of test takers and examiners are observed (see 2.5).

In this study, the analysis of the two primary assessment criteria FC and PD only described three dimensions of conceptual properties, which may also be further expanded if a longer examination task is used so a wider range of concepts that denote the assessment criteria can be elicited. Nevertheless, the three-dimensional aspects of the assessment criteria, though may be simple, are grounded on the examiners' verbal reports which reflect their practical considerations on how they cope with the high cognitive workload when assessing simultaneous interpreting. It appears that a simpler framework is preferred to operationalise the multi-dimensional criteria during the judgement process in a simultaneous interpreting examination.

Due to the complex issues underlying interpreting examinations, however, it would be difficult to investigate all the issues thoroughly and comprehensively in a single study. Therefore, this research study has focused on the initial stage of exploring the intricate core issues of interpreting examinations (see 1.2 and 1.3) by using an examination task with a manageable time length. As discussed in the previous sections, the findings based on the stimulus materials for the examiners has led to valuable understanding and led to the formulation of a conceptual model for interpreting examinations, which may be used as a guide for further studies on relevant topics related to the interpreting examinations.

8.3 The next step for developing the interpreting examinations

As reviewed in Chapter 2, the design and administration of the interpreting examinations in many higher education institutions still heavily rely on the professional experience of staff members, often with no basis of empirical studies for the test items and test procedures (Liu et al., 2008: 35), and the test designs have been described as "intuitive" (Campbell & Hale, 2003: 211). This lack of empirical base has raised serious concerns about the reliability and validity of the interpreting examinations because test constructs and assessment criteria arguably require clear definitions and descriptions. In their introduction to *Testing and Assessment in Translation and Interpreting Studies* (2009), Angelelli and Jacobson 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 (Angelelli & Jacobson, 2009: 21).

Despite the challenges, nevertheless, it is necessary for the community of interpreters, especially those who also teach and research, to acquire the "particular kind of expertise" to define a target construct of the interpreting examinations. As the social consequences of interpreting examinations become more evident, interpreting examinations have to be more carefully scrutinised in the validation process. Sawyer urged that "if validation is a rhetorical art, it is one at which the community of interpreter educators should excel" (2004: 235). After all, if test designers and examiners in the community of interpreters "are unable to express their subjective

judgments by objectively measurable standards" (Kalina, 2005: 768), it will be difficult for the interpreting examinations to be truly reliable. Through research as well as learning from the knowledge in the disciplines of language testing, educational assessment, and perhaps psychology/psychometric testing as the interpreting assessment involves more than language-related skills (e.g. affective factors), the community of interpreters, especially interpreter educators, will be able to find a better footing in developing and improving the interpreting examinations.

As reviewed in 2.4, the development of a test usually begins with the writing of test specifications, which includes three main modules: *construct*, *task*, and *assessment* specifications. Writing the test specifications is also a good way to share experiences and build consensus on how an examination can be developed and improved. Therefore, producing test specifications is a logical next step for developing the interpreting examinations. Being the official statement and blueprint of what an examination tests and how the examination tests it, the test specifications are essential in the establishment of the examination's construct validity (Alderson et al., 1995: 9).

This study has explored and understood the construct of interpreting and how they are operationalised by the identified assessment criteria. The data and findings were generated and inferred from the professional interpreters and teachers who participated in this research study so their judgements and assessment criteria may represent a consensus of a group of practitioners, and are valid overall. Based on the empirical results of this study, therefore, an attempt is made in the section below to produce a working document of *construct* specifications for the interpreting examinations.

The writing of test specifications is "a dynamic process of discussion, piloting and information collection through research" (Fulcher, 2003: 129). Due to the limitation of time and space in this study, this document of construct specifications is only one of the three modules to complete a full set of test specifications; nonetheless, it can be used as

a basis to develop the remaining two modules, i.e. task and assessment specifications, which should include more specific details of the content and procedures of the interpreting examinations. The construct specifications document below is intended to serve as a working document for the continuing process of sharing experiences among test designers, researchers, interpreter trainers and examiners, in the development of the interpreting examinations.

Construct specifications: a proficiency test of simultaneous interpreting[82]

Construct definition

This test assesses the examinees' proficiency in simultaneous interpreting. The examinees should be able to listen to the source speech and simultaneously render the messages faithfully into a target language, conveying them in a way that is clear and coherent for the audience to understand.

The ability to interpret simultaneously is reflected in the following aspects:

- *Message equivalence,* i.e. the ability to convey the messages with faithfulness and meaning to the intent of the original.
- *Language knowledge* that enables an interpreter to correctly understand the speaker's messages and intent in the source language, and to convey them in the target language clearly, coherently, and appropriately according to the context of the speech.
- *Strategic competence*, i.e. the ability to manage linguistic skills and interpreting strategies in a multi-tasking manner to receive, process, and deliver the messages simultaneously.
- *Background knowledge* in the subject area of the speech, including terminology.
- *Influential schemata* that enables an interpreter to cope with high-stress and high-cognitive working conditions in simultaneous interpreting.

Assessment context

The aim of the test is to evaluate student interpreters' achievement in training, and to diagnose their strengths and weaknesses in simultaneous interpreting.

The test is given to student interpreters who have been studying simultaneous interpreting at postgraduate level for one semester. They have had two weekly

[82] This draft of construct specifications was written by following the instructions on writing construct specifications in Luoma's *Assessing Speaking* (2004: 118-121).

lessons and regular self-practice sessions for four months.

The test results are used by both the students and the teacher. The students use the test results as feedbacks on their interpreting performances; the teachers use the test results to evaluate the learning achievements of student interpreters and the effectiveness of the syllabus.

The test scores will also be used to determine whether or not the student interpreters are suitable to be recommended to continue studying simultaneous interpreting in the next stage of the study course.

Description of test

The test task simulates the real-world situation in a conference, and the examinees need to convey the messages across two languages in real time as the speaker speaks to the audience (i.e. the examiners). The test task can be performed by a live speaker, or use the recording of a live speech. The length of the speech task is edited to be ten minutes long with a brief audio instruction to the test task. For background knowledge preparation, the examinees are normally informed about the subject area of the speech task a week before the examination.

The test takes place in a classroom that is equipped with simultaneous interpreting system. The examinees listen to the speech recording in an interpreting booth to perform the task. The examinees' performances are monitored in real time by a panel of two to three interpreter examiners who teach on the same course, but not necessarily teaching the same classes. After each performance, the mark is discussed and agreed onsite by the examination panel.

Construct in context

This is a test of interpreter-mediated communication in a conference where more than one language is used. The flow of information in the speech is unidirectional from a speaker to a group of audience, and the interpretation is provided in real time. The constructs in simultaneous interpreting are assessed by using two main criteria: (1) Fidelity and Completeness, and (2) Presentation and Delivery.

Fidelity and Completeness compares the interpretation with the source speech messages. The criterion is operationalised in three areas: content accuracy, speaker intention, and contextual consistency. The higher the agreement of the three areas is between the interpretation and the source speech, the better the interpreting performance. This criterion is normally the first to be considered and weighted more than the other criterion for its achieving the goal of conveying the speaker's messages and intent, i.e. measuring the construct of *message equivalence*.

Presentation and Delivery observes the way the messages are delivered in the spoken target language. The examiners mainly look at three aspects of the interpretation delivered by the examinees: acoustic effect (i.e. voice texture), idiomatic use of word/phrase, and cohesiveness of the information flow. The more natural and coherent the interpretation is delivered in the target spoken language, the better quality of the interpretation for its user friendliness to the audience. The construct of *language knowledge* is measured mainly by using this criterion.

The *strategic competence* of interpreting is assessed implicitly, so is the examinees' *background knowledge* and listening comprehension of the source speech. The examiners need to diagnose and judge if any omissions of details in the interpretation are good implementation of the condensation norm as an interpreting strategy, or if such omissions and errors are due to problems in listening comprehension, which may be resulted from poor multi-tasking ability or lack of background knowledge support, or if they are due to inadequate speaking competence of the target language. Coupled with the two main criteria, the Ear-Voice Span (EVS) is an observable parameter for reference when making these diagnoses or judgements.

Some *influential schemata* also need to be taken into account in the judgement process, such as the affective factors and personal aptitudes of the examinees, which may be reflected in the Presentation and Delivery criterion, and the physical settings of the examinations, which need to be considered in the design stage of the examination task and the administration of the examination.

Relationship to theoretical models

The construct of interpreting should be perceived and measured within the multi-perspective quality framework of interpreting (Pöchhacker, 2001; Kalina, 2002, 2005), which involves at least four different roles in the examination of the communication act: the speaker, the interpreter, the audience and the examiner. The multi-dimensional quality of interpreting is manifested in Pöchhacker's (2001) quality standards model for interpreting with the accurate rendition of the source speech at the core.

Given that the focus of the test is on interpreter-mediated communication, which requires the use of languages, the construct of simultaneous interpreting can also be related to Bachman and Palmer's (1996) communicative language ability model. The interpreter, i.e. the *speaker* in the language model, is driven by the goal of conveying the messages in the source speech to the target audience across different languages by utilising various language and knowledge resources, and interpreting

skills. At the same time, some affective factors that relate to the interpreters' personal experiences and the physical settings in which they work may also influence the interpreters' performances.

The aspect of *strategic competence* in the construct definition of the IE model can be related to Gile's (1995) Effort Models in Interpretation, which explains how the various efforts are balanced to simultaneously perform different language-related cognitive tasks in simultaneous interpreting.

The conceptual model of the interpreting examinations in this thesis integrates these models to explain the complex relations between various components in the construct of simultaneous interpreting.

The above working document of construct specifications is by no means a final version. There are still many knowledge gaps to be filled (see 8.4 below). The design and development of effective assessment tools and test items to measure the construct of interpreting are all important jobs to be carried out. A full set of test specifications requires the other two modules: assessment and task specifications, which may be developed based on the construct specifications.

8.4 Suggestions for future studies

There are areas that are in need of further studies to secure a sound footing for the interpreting examinations. These areas can be identified and listed by using the four different roles in the IE model. The usefulness of study methodology is also briefly discussed at the end of this section.

8.4.1 The examination task – the speaker

This study only used a three-minute speech task as the test item, and did not use a rating scale in the examination simulation. Although useful findings were generated, more questions need to be asked and answered. For example, to what extent one type of speech can test and generalise the interpreting ability of a student interpreter to other speech types? What rating scales are practical to be used for the examiners as well as for the test users? Are there objective and practical parameters to determine the difficulty levels of the speech tasks? Answering these questions will help produce and develop the task and assessment specifications, and will also help improve the internal consistency and test stability of the interpreting examinations (also see 2.3.3.b and 2.3.3.c).

8.4.2 The interpreter

Little is known about the *influential schemata* that affect the interpreter examinees, which need to be investigated. So far, the personal aptitudes have been studied mainly for the admission examinations, but the study of affective factors in relation to the interpreting assessment is almost none-existent. These influential factors also relate to

the design of examination tasks, such as the physical settings in which the examinations take place. So the consideration here is also related to the examination task, but with a focus on the interaction between the interpreter examinees and the examination tasks. For example, is there a difference in the examination results between using a live speaker and a speech recording (audio and/or video)? Would the examinees interact differently to such two different formats of examination tasks even when the content of the speech is the same? Would the examinees' performances, or the examination results, be influenced by the two different marking approaches – live panel examination marking and post-examination marking from recording? How can the administration of the interpreting examinations be standardised so that the examinees experience similar procedures and conditions when taking the interpreting examinations? In order to remain a valid test, to what extent do the interpreting examinations need to simulate the real-life situations for the interpreter examinees to perform the tasks? These questions will have long-lasting effects on the reliability and validity of the interpreting examination, and on the continuing professional development of interpreting.

8.4.3 The examiner

This study has identified various factors that affect the examiners' assessment behaviours. For example, the *examiner attention* is limited due to high level of cognitive workload in assessment so the examiners resort to holistic marking by impression. Practical assessment tools should be developed to help the examiners reduce their cognitive workload, such as using a speech script for note-taking, or an assessor-oriented rating scale for guiding the marking (Alderson et al., 1995; Fulcher, 2003: 89). The *examiner bias* also plays a role in the judgement process. How does the assessor-audience duel role affect the examiners' judgement? And finally, how do a group of

examiners interact with each other in a jury panel for the interpreting examinations? Studying these issues is useful for designing examiner trainings to help alleviate concerns of reliability due to subjective judgement.

The list of suggestions above for future studies is not exhaustive. Just as mentioned in Chapter 7, the IE model is represented as a balanced system, which is essentially maintained by a two-dimensional tension: the *criteria dimension* and the *behaviour dimension*. To achieve a reliable and valid design of the interpreting examinations, most if not all components in the system need to be carefully checked and balanced.

8.4.4 Some reflections on the study methodology

For a study that is exploratory in nature, the study approach should be as open-ended as possible so that a wider range of data may be collected for analysis, because the researcher is venturing into an unknown field and does not know what may be relevant and what may not be. As outlined in Chapter 1, this research study has taken a multi-strategy approach to maximise the scope of data to be collected for analysis. This approach proved to be useful as it gathered both quantitative and qualitative data for cross-examination. The quantitative methods (i.e. paired comparison and some common statistical analysis) helped establish a framework for analysing the qualitative data, i.e. the interview comments, which in return offered rich information to understand the complex relationships of various construct components of the interpreting examinations that cannot be explained simply by analysing the quantitative data. Using the IE model (Figure 8-1) as a guide, future studies in the field of interpreting assessment may focus on certain key areas, such as those suggested above, by taking study approaches that are more tightly-controlled and hypothesis-testing rather than hypothesis-generating.

8.5 Concluding remark

It has been a long and rewarding research journey to explore the workings of simultaneous interpreting assessment within the educational context. This research study provides an empirical base for raising the awareness of how we assess student interpreters, and for designing and developing future interpreting examinations. It has given a clearer explication of the problems in assessing interpreting performances, and a working model to view the intricate relations of the test constructs, i.e. the IE model; hence, it is hoped that there will be less uncontrolled extraneous variables, or overlooked variables, for conducting future research studies on the test design of the interpreting examinations. In other words, the IE model can be used as a conceptual map for guiding future investigations in the field of the interpreting assessment.

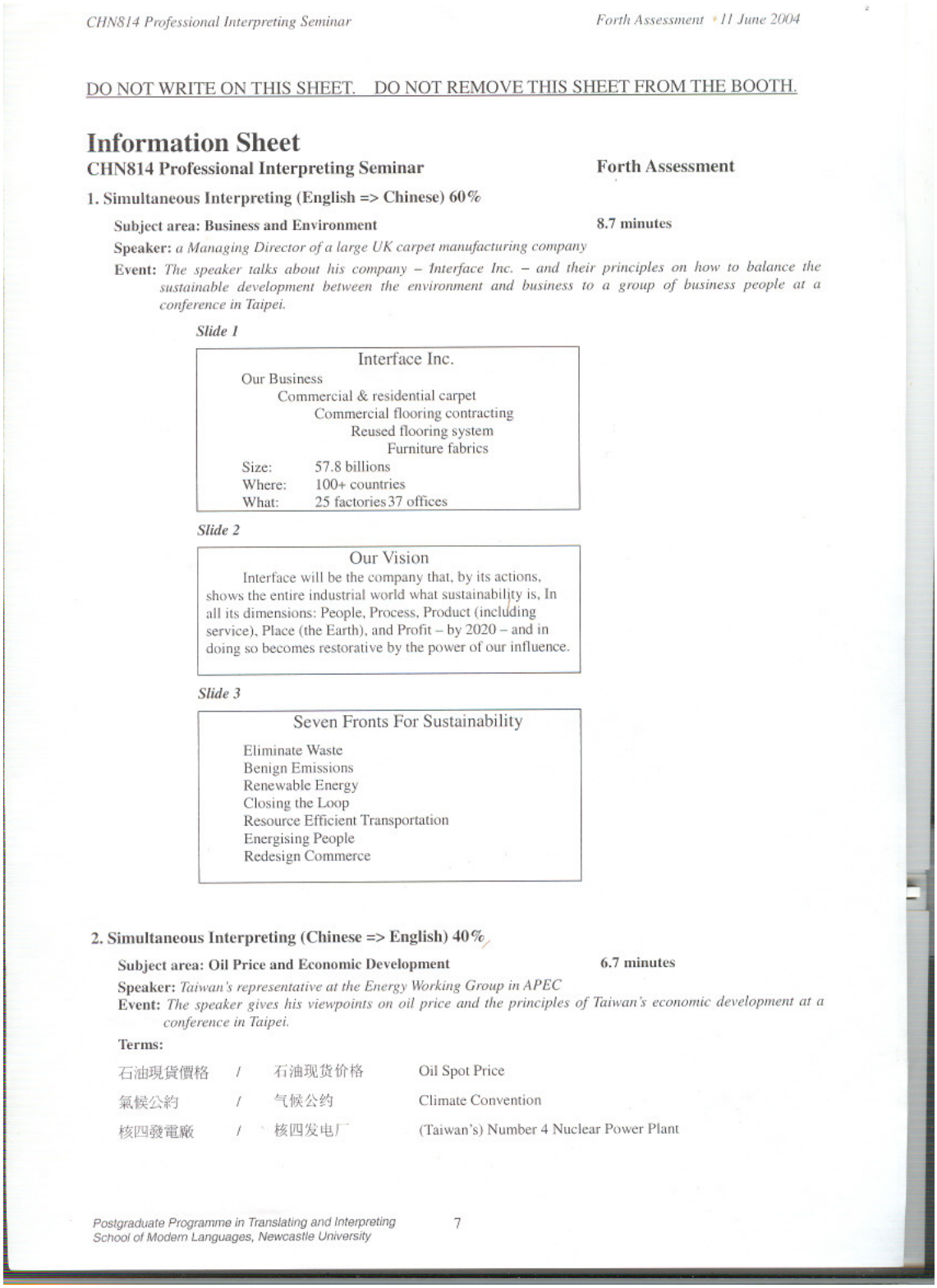
This research journey that the interpreter teacher in the *Prelude* started has come to an end, but new journeys for more explorations have just begun.

* * *

APPENDICES

Appendix A: Examination task

1. Information sheet given to students in the booth



CHN814 Professional Interpreting Seminar *Forth Assessment • 11 June 2004*

DO NOT WRITE ON THIS SHEET. DO NOT REMOVE THIS SHEET FROM THE BOOTH.

Information Sheet

CHN814 Professional Interpreting Seminar **Forth Assessment**

1. Simultaneous Interpreting (English => Chinese) 60%

Subject area: Business and Environment **8.7 minutes**

Speaker: *a Managing Director of a large UK carpet manufacturing company*

Event: *The speaker talks about his company – Interface Inc. – and their principles on how to balance the sustainable development between the environment and business to a group of business people at a conference in Taipei.*

Slide 1

Interface Inc.

Our Business
Commercial & residential carpet
Commercial flooring contracting
Reused flooring system
Furniture fabrics

Size: 57.8 billions
Where: 100+ countries
What: 25 factories 37 offices

Slide 2

Our Vision

Interface will be the company that, by its actions, shows the entire industrial world what sustainability is, In all its dimensions: People, Process, Product (including service), Place (the Earth), and Profit – by 2020 – and in doing so becomes restorative by the power of our influence.

Slide 3

Seven Fronts For Sustainability

Eliminate Waste
Benign Emissions
Renewable Energy
Closing the Loop
Resource Efficient Transportation
Energising People
Redesign Commerce

2. Simultaneous Interpreting (Chinese => English) 40%

Subject area: Oil Price and Economic Development **6.7 minutes**

Speaker: *Taiwan's representative at the Energy Working Group in APEC*

Event: *The speaker gives his viewpoints on oil price and the principles of Taiwan's economic development at a conference in Taipei.*

Terms:

石油現貨價格	/	石油现货价格	Oil Spot Price
氣候公約	/	气候公约	Climate Convention
核四發電廠	/	核四发电厂	(Taiwan's) Number 4 Nuclear Power Plant

Postgraduate Programme in Translating and Interpreting
School of Modern Languages, Newcastle University 7

2. Speech script for the examiners (first 3 minutes used in this study)

Interpreter Code: _____ **Rater Code:**_____ **Date:** _____

*Students will listen to the following speech in **English** about* Business and Environment *simultaneously interpret the message **into Chinese**, which will be recorded on tape for evaluation. The Speaker is a Managing Director of a large UK carpet manufacturing company. He talks about his company – Interface, Inc. – and his principles on how to balance the sustainable development between the environment and business to a group of business people at a conference in Taipei.*

_____ Thank you. It's a quite interesting parallel with business that conducting a large number of people in business and getting all to move in the same direction and with all the egos those individuals have is pretty similar to conducting an orchestra.

_____ The only difference I usually find with an orchestra is the time at the front is quite clear who is in charge; when you are in business, you can't always be at the front and therefore people sometimes forget who is in charge. So there is some parallel to be drawn.

_____ Chairman Chen, President Dickson, distinguished guests: it is an honour to be here from the UK on my first visit to Taiwan.

_____ It's very much a fleeting visit, I am actually going to spend more time in the air travelling here and getting home than I actually will be on the ground. So hopefully next time I'm in Taiwan I should be able to spend much more time here.

_____ I am going to initially in my talk mention briefly who Interface are, what we do, then I'm going to tell you about the journey that we are on, and then I'd like to tell you, towards the end, what we have learned on our journey so far.

_____ We have some interesting things to learn on that journey, some of which are very positive, and some of which are not so positive.

_____ But I'll try and be very honest about where we are, and where we are going as I go through my talk.

_____ But briefly, my job is to travel around as much as I possibly can, talking to people about sustainability, not necessarily about Interface, but talking to people about sustainability.

_____ Because one of the things that will be a recurrent point in my talk is that we have to get to as many people as possible and educate them into the issues connected to sustainability.

_____ And without educating people to what is happening and what may be done about it, we will not make the progress we need to.

_____ So, if we can turn on the next slide, please.

Slide 1

Interface Inc.

Our Business
Commercial & residential carpet
Commercial flooring contracting
Reused flooring system
Furniture fabrics

Size:	57.8 billions
Where:	100+ countries
What:	25 factories 37 offices

_____ We will quickly talk about who Interface are. We're the largest supplier of commercial floor coverings in the world. Somewhere between 35 and 40% of the global market is ours.

_____ And we are also the largest supplier of commercial upholstery fabrics in the world; again, a similar sort of figure globally, but in some parts of the world, as much as 90% of the market.

_____ We clearly have a lot of work to do in Taiwan, because today you are not walking on our floor covering and you are not sitting on our fabric, so it is clear we have quite a lot of work to do here in Taiwan.

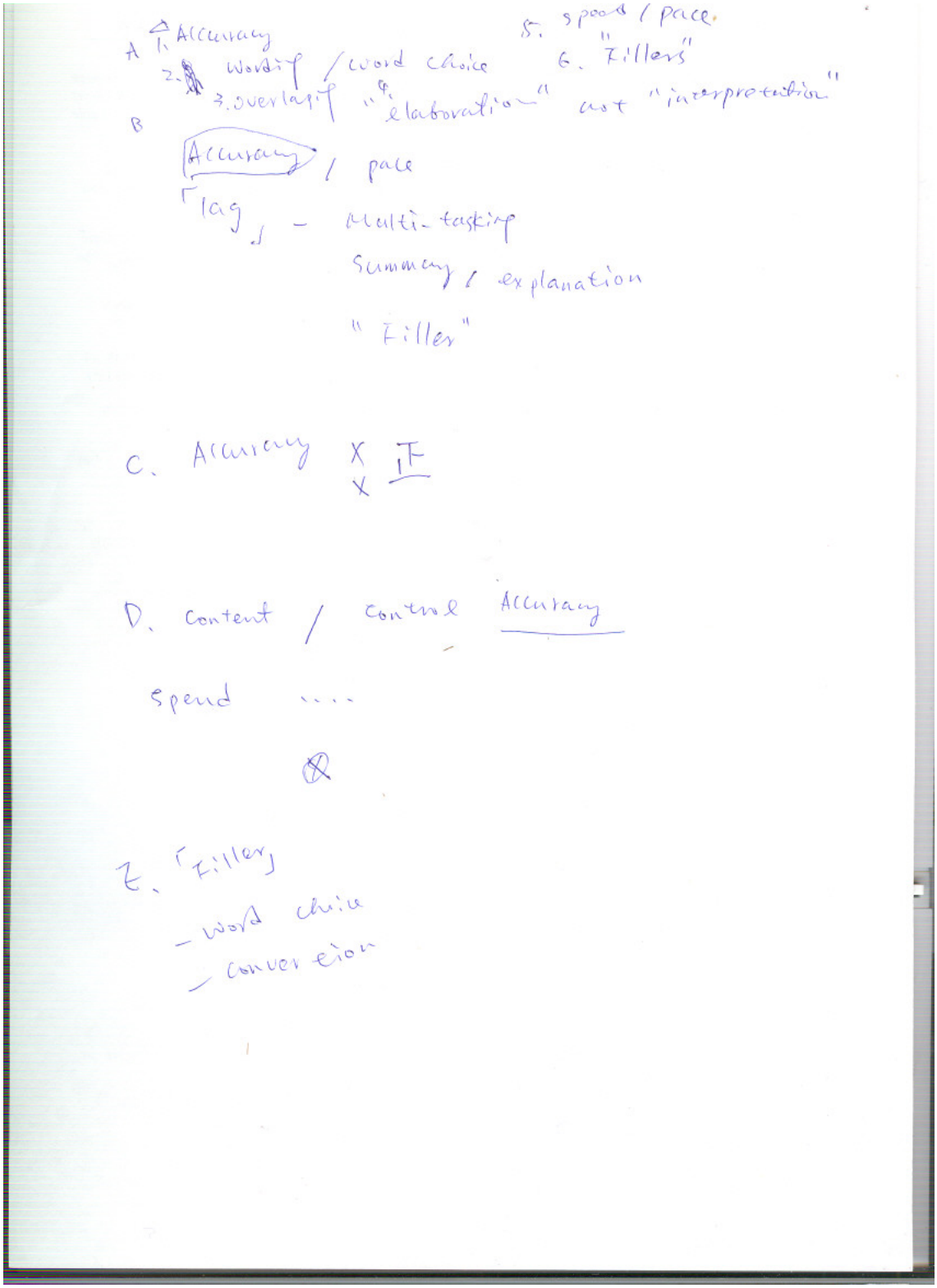
_____ Our turnover last year was 1.1 billion US dollars, which in the interior's market that we belong to, makes us huge.

_____ In Wall Street terms as a Wall-Street-declared company, that makes us very small.

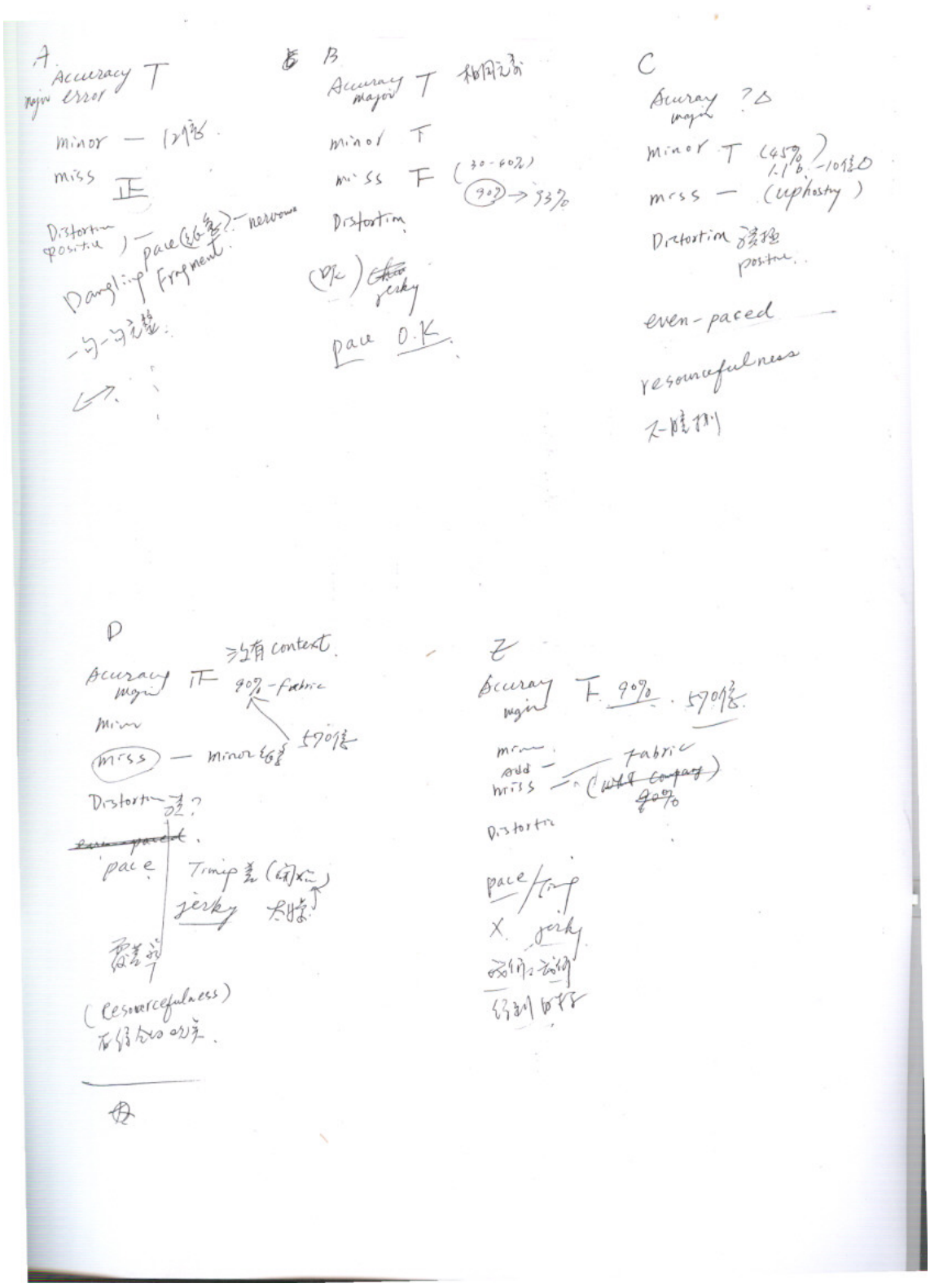
_____ So we have sort of two different opinions about our position to work with.

Appendix B: Samples of examiners' notes

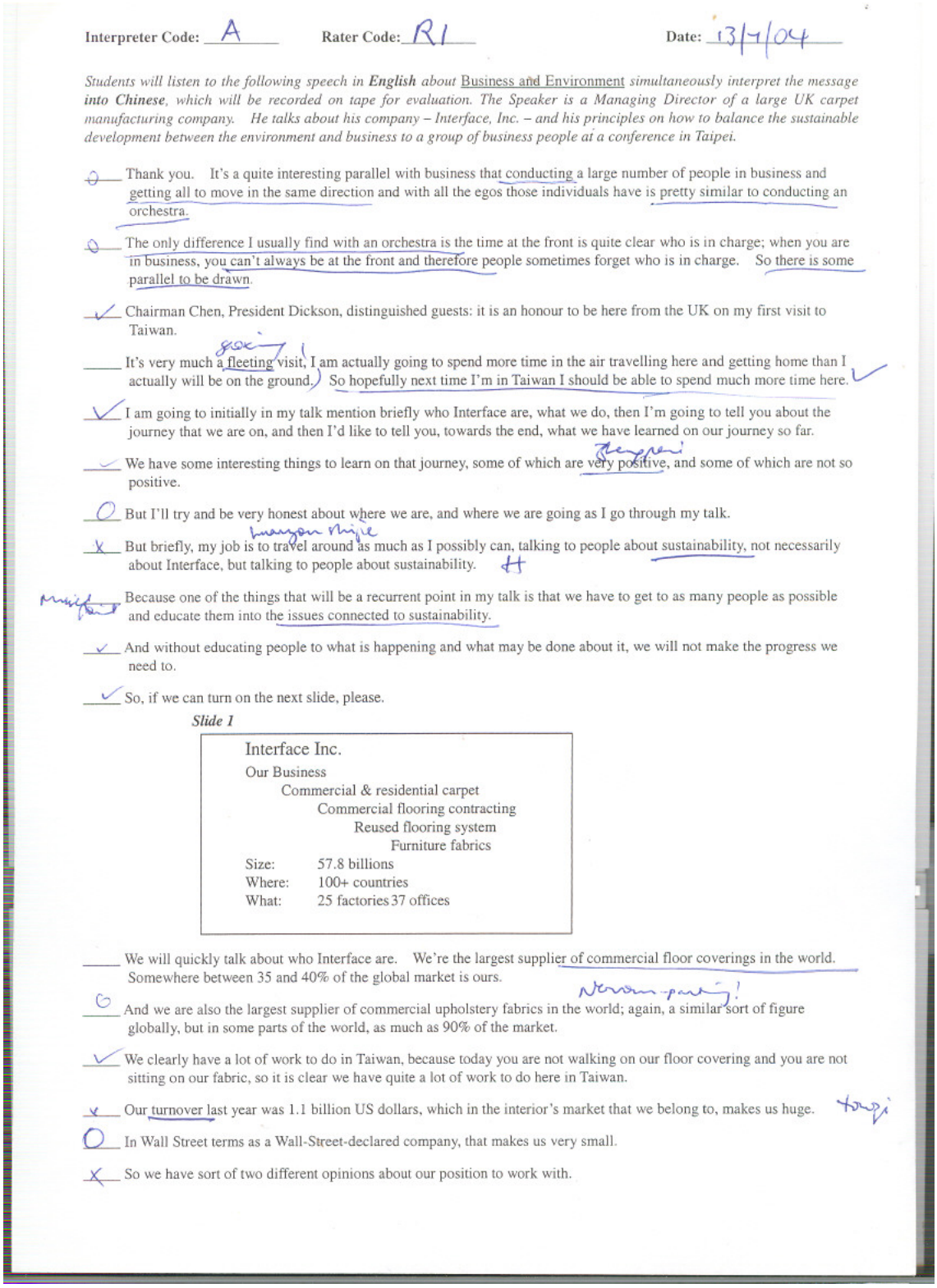
Sample 1: Taking notes without referring to speech script



Sample 2: Taking notes without referring to speech script



Sample 3: Taking notes with speech script



Interpreter Code: A Rater Code: R1 Date: 13/4/04

*Students will listen to the following speech in **English** about Business and Environment simultaneously interpret the message **into Chinese**, which will be recorded on tape for evaluation. The Speaker is a Managing Director of a large UK carpet manufacturing company. He talks about his company – Interface, Inc. – and his principles on how to balance the sustainable development between the environment and business to a group of business people at a conference in Taipei.*

Thank you. It's a quite interesting parallel with business that conducting a large number of people in business and getting all to move in the same direction and with all the egos those individuals have is pretty similar to conducting an orchestra.

The only difference I usually find with an orchestra is the time at the front is quite clear who is in charge; when you are in business, you can't always be at the front and therefore people sometimes forget who is in charge. So there is some parallel to be drawn.

Chairman Chen, President Dickson, distinguished guests: it is an honour to be here from the UK on my first visit to Taiwan.

It's very much a fleeting visit, I am actually going to spend more time in the air travelling here and getting home than I actually will be on the ground. So hopefully next time I'm in Taiwan I should be able to spend much more time here.

I am going to initially in my talk mention briefly who Interface are, what we do, then I'm going to tell you about the journey that we are on, and then I'd like to tell you, towards the end, what we have learned on our journey so far.

We have some interesting things to learn on that journey, some of which are very positive, and some of which are not so positive.

But I'll try and be very honest about where we are, and where we are going as I go through my talk.

But briefly, my job is to travel around as much as I possibly can, talking to people about sustainability, not necessarily about Interface, but talking to people about sustainability.

Because one of the things that will be a recurrent point in my talk is that we have to get to as many people as possible and educate them into the issues connected to sustainability.

And without educating people to what is happening and what may be done about it, we will not make the progress we need to.

So, if we can turn on the next slide, please.

Slide 1

Interface Inc.	
Our Business	
Commercial & residential carpet	
Commercial flooring contracting	
Reused flooring system	
Furniture fabrics	
Size:	57.8 billions
Where:	100+ countries
What:	25 factories 37 offices

We will quickly talk about who Interface are. We're the largest supplier of commercial floor coverings in the world. Somewhere between 35 and 40% of the global market is ours.

And we are also the largest supplier of commercial upholstery fabrics in the world; again, a similar sort of figure globally, but in some parts of the world, as much as 90% of the market.

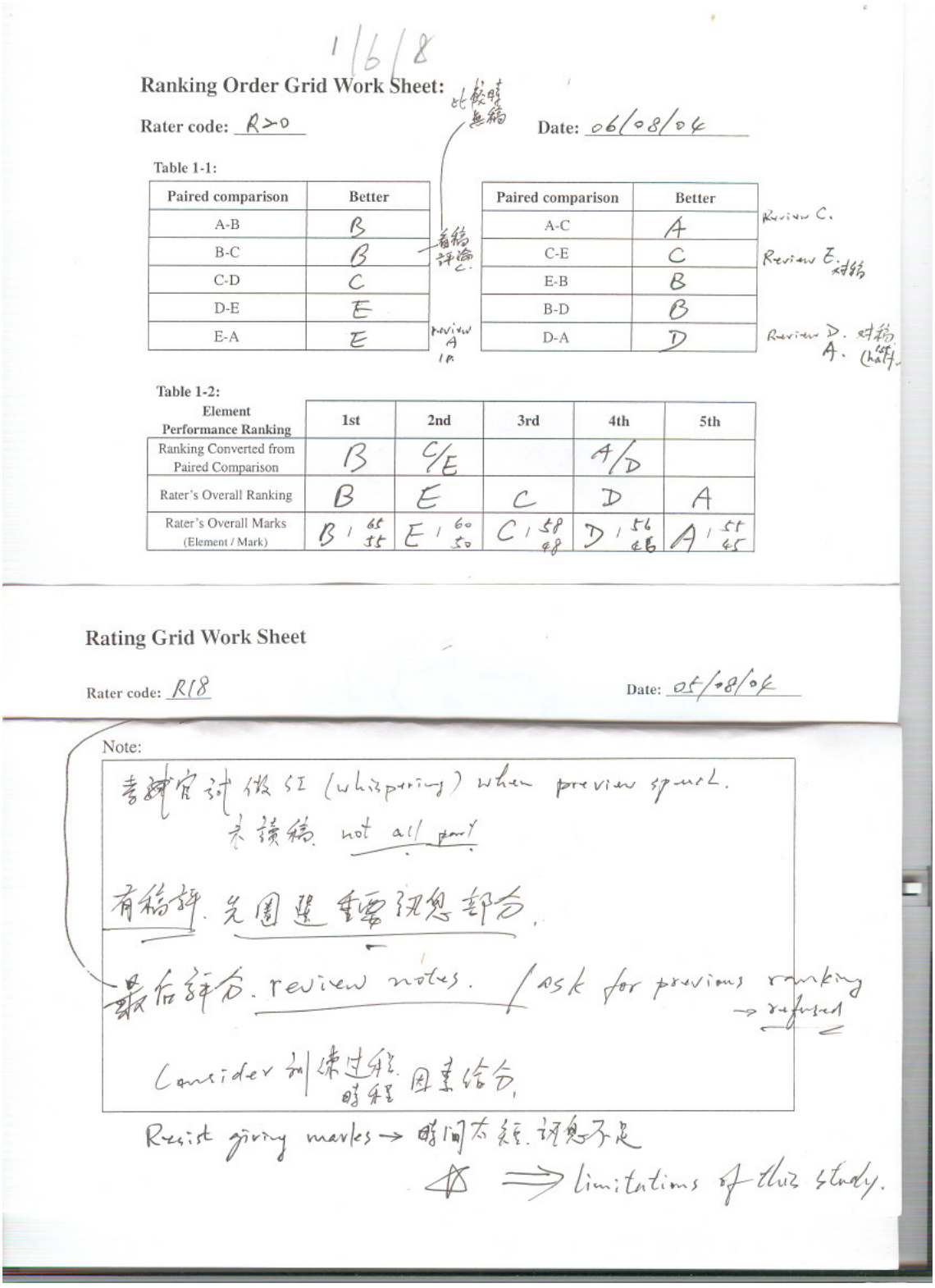
We clearly have a lot of work to do in Taiwan, because today you are not walking on our floor covering and you are not sitting on our fabric, so it is clear we have quite a lot of work to do here in Taiwan.

Our turnover last year was 1.1 billion US dollars, which in the interior's market that we belong to, makes us huge.

In Wall Street terms as a Wall-Street-declared company, that makes us very small.

So we have sort of two different opinions about our position to work with.

Appendix C: Sample of the researcher's field notes in worksheets



1/6/8

Ranking Order Grid Work Sheet:

Rater code: R20 Date: 06/08/04

Table 1-1:

Paired comparison	Better
A-B	B
B-C	B
C-D	C
D-E	E
E-A	E

Paired comparison	Better
A-C	A
C-E	C
E-B	B
B-D	B
D-A	D

Table 1-2:

Element Performance Ranking	1st	2nd	3rd	4th	5th
Ranking Converted from Paired Comparison	B	C/E		A/D	
Rater's Overall Ranking	B	E	C	D	A
Rater's Overall Marks (Element / Mark)	B / 65 55	E / 60 50	C / 58 48	D / 56 46	A / 55 45

Rating Grid Work Sheet

Rater code: R18 Date: 04/08/04

Note:

書紙官討做 SI (whispering) when preview speak.
未讀稿 not all part

有稿評. 先圍選重要視點部分.

最後評分. review notes. / ask for previous ranking → refused

Consider 訓練過程 時程 因素給分.

Resist giving marks → 時間太短. 訪問不足
⇒ limitations of this study.

Appendix D: Paired comparison winners according to cluster membership

A: Ally B: Beth C: Cherry D: Daisy E: Eileen =: draw

Cluster examiners		Paired comparison winners									
		1. A-B	**2. B-C**	**3. C-D**	**4. D-E**	**5. E-A**	**6. A-C**	**7. C-E**	**8. E-B**	**9. B-D**	**10. D-A**
PCC1	**R1**	B	C	C	E	E	C	C	**E**	B	D
	R30	B	C	C	E	E	C	C	**E**	B	D
	R28	**A**	C	C	E	**A**	C	C	**E**	B	**A**
	R16	B	C	C	E	E	C	C	**E**	**D**	D
	R15	B	C	C	E	E	C	C	B	B	D
	R27	B	C	C	E	E	C	C	B	B	D
	R3	B	C	C	E	E	C	C	B	B	D
	R7	B	C	C	E	E	C	C	B	B	D
	R11	B	C	C	**D**	E	C	C	B	B	D
	R29	B	C	C	=	E	C	C	=	B	D
PCC2	**R4**	B	**C**	C	E	**A**	C	C	B	B	A
	R21	B	**C**	C	E	**A**	C	C	B	B	A
	R8	B	**C**	C	E	E	C	C	B	B	A
	R13	B	**C**	C	E	E	C	C	B	B	A
	R5	B	B	C	E	E	C	C	B	B	A
	R23	B	B	C	E	E	C	C	B	B	A
	R6	B	B	C	E	E	C	C	B	B	**D**
	R25	B	B	C	E	E	C	C	B	B	**D**
	R20	B	B	**C**	E	E	**A**	C	B	B	**D**
	R2	B	B	C	E	E	C	**E**	**E**	B	A
PCC3	**R9**	B	**B**	C	D	A	**A**	C	B	**D**	A
	R10	B	**B**	C	D	A	**A**	C	B	B	A
	R12	B	**C**	C	D	A	**C**	C	B	B	A
	R17	**A**	**C**	C	D	A	**C**	C	B	B	A
PCC4	**R18**	B	C	**D**	D	**E**	C	C	B	D	D
	R26	B	C	**C**	D	**A**	C	C	B	D	D
PCC5	**R22**	B	B	**C**	D	E	C	C	B	B	D
	R24	B	B	**D**	D	E	C	C	B	B	D
	R19	B	B	**C**	D	**A**	C	C	B	B	D
	R14	B	B	**D**	D	E	C	**E**	B	B	D

Appendix E: z-scores – PC, OJ and OM

Paired Ranking z-scores

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20	R21	R22	R23	R24	R25	R26	R27	R28	R29	R30
A	-1.26	-0.63	-1.26	0.00	-0.63	-1.26	-1.26	-0.63	0.96	0.63	-1.26	0.00	-0.63	-1.26	-1.26	-1.26	0.63	-1.26	-0.63	-0.96	0.00	-1.26	-0.35	-1.26	-1.26	-0.63	-1.26	-0.63	-1.41	-1.26
B	0.00	0.63	0.63	0.63	1.26	1.26	0.63	0.63	0.96	1.26	0.63	0.63	0.63	1.26	0.63	-0.63	0.00	0.00	1.26	1.43	0.63	1.26	1.40	1.26	1.26	0.00	0.63	0.00	0.00	0.00
C	1.26	0.00	1.26	1.26	0.63	0.63	1.26	1.26	-0.24	0.00	1.26	1.26	1.26	-0.63	1.26	1.26	1.26	0.63	0.63	0.24	1.26	0.63	0.53	0.00	0.63	1.26	1.26	1.26	1.41	1.26
D	-0.63	-1.26	-0.63	-1.26	-1.26	-0.63	-0.63	-1.26	-0.24	-0.63	0.00	-0.63	-1.26	0.63	-0.63	0.00	-0.63	1.26	0.00	-0.96	-1.26	0.00	-1.23	0.63	-0.63	0.63	-0.63	-1.26	0.00	-0.63
E	0.63	1.26	0.00	-0.63	0.00	0.00	0.00	0.00	-1.43	-1.26	-0.63	-1.26	0.00	0.00	0.00	0.63	-1.26	-0.63	-1.26	0.24	-0.63	-0.63	-0.35	-0.63	0.00	-1.26	0.00	0.63	0.00	0.63

Overall Ranking z-scores

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20	R21	R22	R23	R24	R25	R26	R27	R28	R29	R30
A	-1.26	-0.63	-1.26	0.00	-0.63	-1.26	-1.26	0.00	1.26	0.63	-0.63	0.00	-0.63	-1.26	-1.26	-1.26	0.00	-1.26	-0.63	-1.26	0.00	-1.26	-0.63	-1.26	-1.26	-0.63	-1.26	-0.63	-1.41	-1.26
B	0.63	0.00	0.63	0.63	1.26	0.63	0.63	0.63	1.26	0.63	1.26	0.63	1.26	0.63	-0.63	0.63	0.00	1.26	1.26	0.63	1.26	0.63	1.26	1.26	0.00	0.63	0.00	0.00	0.00	-0.63
C	1.26	0.63	1.26	1.26	0.63	1.26	1.26	1.26	-0.63	0.00	1.26	0.63	1.26	0.00	1.26	1.26	1.26	0.63	0.00	1.26	0.63	1.26	0.63	0.63	1.26	1.26	1.26	1.41	1.26	
D	-0.63	-1.26	-0.63	-1.26	-1.26	-0.63	-0.63	-1.26	0.00	-0.63	0.00	-0.63	-1.26	0.63	0.00	0.00	-0.63	0.63	0.00	-0.63	-0.63	0.00	-1.26	0.00	-0.63	0.63	0.00	-1.26	0.00	0.00
E	0.00	1.26	0.00	-0.63	0.00	0.00	0.00	-0.63	-1.26	-1.26	-1.26	0.00	-0.63	-0.63	0.63	-1.26	-0.63	-1.26	0.63	-1.26	-0.63	0.00	-0.63	0.00	-1.26	-0.63	0.63	0.00	0.63	

Student Marks - z-scores

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20	R21	R22	R23	R24	R25	R26	R27	R28	R29	R30
A	-1.07	-0.26	-0.84	-0.36	-0.13	-0.89	-1.40	0.07	0.67	0.85	-1.10	-1.73	-0.70	-1.38	-1.43	-1.76	0.17	-1.16	-0.59	-0.96	-0.07	-1.47	-0.41	-1.74	-1.07	-0.05	-1.05	0.04	-1.59	-1.56
B	-0.70	0.26	0.96	0.61	0.88	1.11	-0.18	0.79	1.12	1.24	0.45	0.74	1.04	1.41	0.40	0.42	0.77	-0.28	1.26	1.56	0.95	1.12	0.66	0.75	1.38	0.63	0.92	0.04	0.13	0.19
C	1.52	0.53	1.16	1.16	1.16	1.02	1.34	1.15	0.06	-0.48	1.48	0.05	1.04	-0.27	1.32	0.66	1.13	1.13	0.52	-0.20	0.95	0.69	1.09	0.44	0.49	1.31	1.20	1.14	1.19	1.13
D	0.04	-1.58	-0.94	-1.47	-0.87	-0.35	-0.18	-1.01	-0.40	-0.48	-0.24	0.60	-1.13	0.07	-0.15	0.17	-1.03	0.95	0.15	-0.71	-0.41	-0.03	0.13	0.44	0.04	-0.74	-0.68	-1.60	0.13	-0.22
E	0.22	1.05	-0.34	0.06	-1.05	-0.89	0.43	-1.01	-1.46	-1.14	-0.58	0.33	-0.26	0.18	-0.15	0.51	-1.03	-0.63	-1.33	0.30	-1.42	-0.32	-1.48	0.12	-0.85	-1.15	-0.39	0.37	0.13	0.46

Appendix F: ANOVA statistics

Sample 1: Reliability test of the three assessment methods

Descriptives

		N	Mean	Std. Deviation	Std. Error
Zscore(pc)	A	30	-.8344233	.78969553	.14417802
	B	30	.7448525	.61404009	.11210787
	C	30	.9334227	.62517163	.11414020
	D	30	-.5515680	.75851094	.13848452
	E	30	-.2922839	.75661418	.13813822
	Total	150	.0000000	1.00000000	.08164966
Zscore(oj)	A	30	-.8249579	.74471317	.13596540
	B	30	.7071068	.61588176	.11244411
	C	30	1.0370899	.57933771	.10577211
	D	30	-.4949747	.67338568	.12294284
	E	30	-.4242641	.80086160	.14621666
	Total	150	.0000000	1.00000000	.08164966
Zscore(om)	A	30	-.6840636	1.06330613	.19413225
	B	30	.5842139	.67611986	.12344203
	C	30	.7423609	.77204410	.14095532
	D	30	-.3739713	.78279215	.14291764
	E	30	-.2685399	.84547196	.15436136
	Total	150	.0000000	1.00000000	.08164966

Descriptives

		95% Confidence Interval for Mean			
		Lower Bound	Upper Bound	Minimum	Maximum
Zscore(pc)	A	-1.1293005	-.5395462	-1.42371	1.40485
	B	.5155661	.9741388	-.71657	1.40485
	C	.6999798	1.1668656	-.71657	1.40485
	D	-.8348006	-.2683353	-1.42371	1.40485
	E	-.5748083	-.0097595	-1.42371	1.40485
	Total	-.1613408	.1613408	-1.42371	1.40485
Zscore(oj)	A	-1.1030384	-.5468774	-1.41421	1.41421
	B	.4771328	.9370808	-.70711	1.41421
	C	.8207617	1.2534182	-.70711	1.41421
	D	-.7464211	-.2435284	-1.41421	.70711
	E	-.7233107	-.1252174	-1.41421	1.41421
	Total	-.1613408	.1613408	-1.41421	1.41421
Zscore(om)	A	-1.0811086	-.2870185	-4.69356	.60902
	B	.3317466	.8366812	-.97245	1.81838
	C	.4540749	1.0306469	-1.15850	1.81838
	D	-.6662707	-.0816719	-2.83300	1.35324
	E	-.5842443	.0471645	-1.90273	1.81838
	Total	-.1613408	.1613408	-4.69356	1.81838

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Zscore(pc)	Between Groups	75.360	4	18.840	37.097	.000
	Within Groups	73.640	145	.508		
	Total	149.000	149			
Zscore(oj)	Between Groups	80.433	4	20.108	42.524	.000
	Within Groups	68.567	145	.473		
	Total	149.000	149			
Zscore(om)	Between Groups	47.170	4	11.792	16.792	.000
	Within Groups	101.830	145	.702		
	Total	149.000	149			

Sample 2: Reliability test of salient criteria use – PC clusters

Case Processing Summary

		N	%
Cases	Valid	5	100.0
	Excluded[a]	0	.0
	Total	5	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.982	.986	5

Inter-Item Correlation Matrix

	PCC1	PCC2	PCC3	PCC4	PCC5
PCC1	1.000	.987	.917	.961	.977
PCC2	.987	1.000	.954	.904	.967
PCC3	.917	.954	1.000	.790	.917
PCC4	.961	.904	.790	1.000	.941
PCC5	.977	.967	.917	.941	1.000

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance
Inter-Item Correlations	.931	.790	.987	.197	1.250	.003

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
PCC1	79.800	7796.825	.992	.	.971
PCC2	79.800	8210.325	.986	.	.973
PCC3	79.900	7609.300	.911	.	.988
PCC4	79.800	8698.325	.911	.	.984
PCC5	79.900	8471.675	.980	.	.975

ANOVA

		Sum of Squares	df	Mean Square	F	Sig
Within People	Between People	10151.660	4	2537.915		
	Between Items	.060	4	.015	.000	1.000
	Residual	719.240	16	44.953		
	Total	719.300	20	35.965		
	Total	10870.960	24	452.957		

Grand Mean = 19.960

Intraclass Correlation Coefficient

		95% Confidence Interval	
	Intraclass Correlation[a]	Lower Bound	Upper Bound
Single Measures	.917[b]	.739	.990
Average Measures	.982	.934	.998

Two-way random effects model where both people effects and measures effects are random.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

Sample 3: Reliability test of salient criteria use – Beth-Cherry comparison

Case Processing Summary

		N	%
Cases	Valid	5	100.0
	Excluded[a]	0	.0
	Total	5	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.930	.956	5

Inter-Item Correlation Matrix

	PCC1	PCC2	PCC3	PCC4	PCC5
PCC1	1.000	.955	.899	.715	.973
PCC2	.955	1.000	.736	.477	.973
PCC3	.899	.736	1.000	.943	.840
PCC4	.715	.477	.943	1.000	.612
PCC5	.973	.973	.840	.612	1.000

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance
Inter-Item Correlations	.812	.477	.973	.496	2.041	.028

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
PCC1	80.000	13487.500	.950	.	.902
PCC2	80.000	14237.500	.791	.	.925
PCC3	80.000	11400.000	.970	.	.883
PCC4	80.000	10462.500	.738	.	.962
PCC5	80.000	12962.500	.882	.	.905

ANOVA

		Sum of Squares	df	Mean Square	F	Sig
	Between People	15370.000	4	3842.500		
Within People	Between Items	.000	4	.000	.000	1.000
	Residual	4280.000	16	267.500		
	Total	4280.000	20	214.000		
	Total	19650.000	24	818.750		

Grand Mean = 20.000

Intraclass Correlation Coefficient

		95% Confidence Interval	
	Intraclass Correlation[a]	Lower Bound	Upper Bound
Single Measures	.728[b]	.363	.961
Average Measures	.930	.740	.992

Two-way random effects model where both people effects and measures effects are random.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

Appendix G: Sample transcription of examiner comments

(Examiner code is deleted here.)

比較上來說的話，B 做得比 A 好。第一個我覺得 B 講話的速度比較慢，如果就一個完全不懂英文的人來講的話，他只能聽口譯員，那我覺得 B 的速度我比較能聽得懂他在講什麼。就是估且不論內容的正確性，他的速度慢，我不會覺得壓迫性很大。A 的速度很快，聽的人覺得，哇，很緊張。這是第一個，速度。然後我覺得在內容上來講的話，A 給我的感覺，在前半段他大概都只有抓到頭跟尾，那中間幾乎是不見的。他比較像是我們剛開始學口譯時候的那個樣子。因為可能我一開口我就聽不到，所以第一句我一定聽得到，所以我一定做得到第一句，然後當我做的時候，中間就不見了，所以我趕快再聽到最後一句，我又會做到最後一句。那這樣子做起來就變成他沒有辦法，就算他要幺他都沒辦法幺，因為中間不見了。那他做到中段之後我覺得他開始可能比較安心一點，所以內容就開始比較多了。可是跟 B 比較起來的話，他還是不夠。就是好，B 做得比較慢，而且我感覺是 B 好像比較有準備，或者是他的工夫比較深，所以他基本上都可以慢慢地聽聽聽，聽得到，那可能就變成說，就是他都聽得到，所以你會聽到比較多的訊息。那在技巧方面，我覺得有一個是他對於 term 的掌握度比較高。像我們會很清楚知道 sustainability 是永續發展，那他用一個我們很熟的詞或者是他也很熟的詞，他就不需要處理很多。他只要碰到這個字他就是永續發展永續發展。那就變成，我覺得他可以比較懂得怎麼樣去分配他的腦力。那我從他這邊可以得到比較多的訊息，我覺得他的那個 fidelity 就會比較高。那這兩個比較起來，我覺得 B 很明顯地比 A 好很多。(2:36)

*（B 跟 C 做比較）*我覺得這兩個很難比較誰比較好。C 的速度更慢，但是卻沒有，就是他講話速度好慢，可是我就在想說他會不會落掉，可是我後來發現，他當然會有遺漏的部份，可是跟 B 比較起來好像都差不多。就是他們掉的那個，他們對資訊掌握的程度好像都差不多。那如果是這樣子的話，我會覺得 C 給我的感覺是好像講得更從容。而且我發現 C 只要他很清楚掌握到的部份，他做的正確性就很高，但是如果他沒有掌握到的部份，就會出問題，可是誰都一樣，沒有掌握到的部份，都會出問題嘛。但是我覺得他的掌握的程度，如果他有掌握地比較清楚的，那他就會講得很清楚，而且他會考慮到首尾連貫的問題。這一段可能有三句，B 會讓我覺得他三句可能都翻對，可是他只是把每一句忠實地翻出來；那 C 會讓我覺得他甚至好像他的那個腦子轉的更快，他有時間聽完三句之後，然後會調動一下那個，可能要先第一句講一半，然後第二句講一半，再回來講第一句，然後再講第三句，這樣子聽起來會比較通順，我覺得他好像有這樣子的能力。就是不是完全這樣子硬翻。那 B 也會翻對，C 也會翻對，但是 C 顯然多做了一個潤飾的工作。所以我在想，我不知道他們是不是有先看過稿子？*（沒有，這是他們第一次聽。）*如果是這樣子，那我覺得 C 做得比較好。因為我本來懷疑說他是不是有先記得什麼東西，如果沒有的話，那我就認為 C 做得比較好。就是他掌握的訊息，不對，應該這樣講，他在口譯的結果是比較好的，他的，應該怎麼講，他的東西聽起來像是有經過潤飾的感覺，比較能夠首尾連貫。可能對聽眾來講，我也不用花太多時間腦筋去思考，我就是聽你在演講，一個…比較像一個中文的演講。(5:46)

比較 C 跟 D 的話我還是覺得 C 做得比較好。D 的速度聽起來是蠻舒服的，可是我覺得 D 第一個有一點就是，他聽到的訊息並沒有 C 聽到得多。從他講出來的東西你會發現，他很明顯地會掉一些東西，可能四句掉一句，那這是第一個問題。第二個問題我覺得他是不是有點像 over translation，他會冒出很多講者沒有講的東西。比如說他在第一句他就講到永續發展那個東西，可是事實上並沒有啊。就是講者根本還沒有破題，什麼都沒有。然後像，有一個我不知道是不是大陸那邊用語的問題，他說在我的講話裡面，然後用到教訓，我覺得這個是不是有點 over 的可能。如果他們的習慣用語不是這樣的話。最 over 的地方是講到 progress 的時候，他竟然指明的講經濟發展。可是我覺得在這邊好像，其實講者並沒有這麼指明地講我要講經濟發

展，因為他現在講的是他的公司嘛，而不是說整個那個。那到後來好像 D 給我的感覺可能累了吧，後面掉的東西更多。因為像他在介紹 interface 的市場佔有率是，在這個 commercial floor covering 他完全沒有講到，這個 upholstery fabric 也都沒有講到，那變成我只聽到說，我們有 35 到 40 的市佔率，可是是什麼不知道。我覺得後面這邊訊息掉得更嚴重，甚至連數字都會，他後來冒出一個 570 億英磅，可是根本沒有講這個啊。他演講的那個時候牆壁上有打這個嗎？*（他們手上有投影片的資料。）*對對對。這是故意要誤導他們嗎？*（沒有誤導，實際的場合就是這樣子。）*我覺得 C 做得比較好。(8:40)

*（請比較 D 跟 E。）*如果一定要比的話，我覺得 E 比較好。因為 E 在中間這一段的表現上，就是你會很明顯地聽到 E 在這裡講話變得比較慢，然後可能，我想是因為這一段比較簡單一點。就是在講那個旅行的事情。那 E 在這邊講得比較慢，但是他講得比較完整。那 D 是從頭到尾並沒有一個比較可以 mark 出來說，啊，你這段做得不錯的地方，沒有。那 E 當然也不是說那麼好，但是 E 至少在這一段上面，因為 E 一開始根本就是在亂講啊。就是他第一句，第一個 section 他根本不知道在講什麼，然後第二個 section 大概掉了三分之二，然後從這邊開始，打招呼這邊是一個很明顯的分水嶺，因為打招呼大家就有點不太管你在講什麼，反正就是什麼，各位來賓大家好。所以好像輕鬆一點的時候他就開始聽這邊聽得比較清楚，然後就開始發現他開始好轉。然後接下來這段因為是比較簡單的，所以，這邊他一樣有掉，可是我就覺得他掉得還 ok，他這邊表現得比較好。然後這邊開始快，可能有一點信心了我覺得。他就是速度開始變快，但是這邊的東西沒有這麼容易。因為講到市場佔有率，又講到那個市場，這邊同時出現兩組數字嘛，然後又兩組不一樣的東西，然後後面又出現一個數字。到這邊開始又變了，他一樣啊，他也是看這個。我覺得 E 比較好，他至少中間有一段掌握地還不錯。(11:02)

*（比較 E 跟 A）*我覺得 E 比較好。原因是就不好的方面來說的話，他們兩個人都是，就是前面兩段，我剛剛說 E 有在亂掰，那 A 也是一樣。他們同樣都在中間這一段表現得比較好，就是兩個人都是開始好轉。可是 E 給我的感覺比較像在做口譯。因為 A 講話速度比較快，我剛剛再注意去聽了一下，我發現就是他落後得比較多。A 落後比較多，然後他講話又比較快，我說 E 比較像口譯是因為我覺得 A 比較像是我聽了三句話之後，就是我的，我們這樣講好了，A 的記憶可能比 E 好，A 的 short-term memory 可能比 E 強。那我不知道這樣算不算比較不適何做口譯的能力，就是比如說，我仗著我的記憶力比較好，所以我聽三句，再很快的丟出來，所以變得很趕嘛。因為如果我是一句一句來的話，我可以講得比較從容一點，然後讓我覺得 E 就是那種他幺的功力會比較強，因為他是貼著在做的，他比較貼近講者的 tempo，所以我順著你這樣講的話那因為中英文結構的關係，他勢必要去做一點點幺的那個動作，加一點點介繫詞啊，但是我覺得他這樣是比較貼近口譯的作法。A 讓我的感覺就是，如果我的口譯技巧比較不好，但是我發現我的記憶力還不錯的話，我可能會採取的就是說，我一定要聽到完我才知道你在講什麼，那就變成是理解而不是做，就變成是一個完整訊息的理解而不是像口譯這樣，一直接受一直接受，*（同時聽同時講?）*對對，我覺得是這樣。所以我覺得 E 做得比較好，因為他比較貼近所謂的同步口譯。*（那可不可以說，這兩個比較，E 比較好是因為你覺得 E 的技巧比較好？）*對，就做口譯的技巧來講的話，可能如果我是老師的話，我會覺得 E 讓我看到進步。一定是從視譯開始教嘛，這樣加下來的話，我會覺得 E 是一個有在進步的學生。以這個考試來說的話，A 可能還沒有達到我要求學生做的，可能我要求學生就是我要你能夠做同口不要那麼仰賴你的記憶，這 A 可能還差一關。這樣的話我覺得 E 比較好。(14:22)

*(A 跟 C 請比較)*我覺得就是 C 比較好。因為 C 一開始給我的感覺就是我覺得他做得不錯，我就覺得他是不是事先有先看過這個東西嘛。那 A 跟 C 的差別就是，第一個因為 A 的速度太快，然後我甚至覺得那個 voice quality，A 有一點太尖銳。那發音的部份就沒有辦法，因為我覺得他跟我們是不一樣的。我覺得那可能不能評斷。那 A 的聲音比較尖，速度比較快，那更何況在訊息的掌握上的掌握度要比 A 來得高。我記得 C 比較沒有失真的那個問題，A 跟 C 的話，我覺得 C 比較好。(15:43)

*（那E 跟B 呢）*E 跟 B 的話我會覺得 B 還是比 E 好。因為我們剛剛比較 A 跟 E 嘛。*（不要管別的，就是E 跟B）*E 跟 B 的話我覺得 B 是比較好的。因為 B 講話，B 跟 E 講話速度差不多慢，但是 B 掌握到的 information 比較多，比 E 來得好。B 在詞彙的掌握上也是比 E 來得強。其實這可能跟，不曉得，這是不是跟背景有關係。因為如果說他們之前沒有人有準備這個東西的話，我覺得像什麼市場佔有率啊、永續發展啊這種東西感覺上 B 就來得要好多了。對，B 比較好。(16:49)

*（現在請比較 B 跟 D）*B 跟 D，我覺得 B 比較好。他們的速度差不多，那其實在抓到的 information 的多寡我覺得也是差不多，但是 D 我覺得有一個問題就是他有 over translation 的問題。那我想這個不是一件好事。而且平均來講的話，D 在後面出的差錯是比較大的，就是他該講的重點沒有講出來，然後數字什麼的有問題。我覺得 B 比較好。

*（最後一個請比較D 跟A）*我可以聽一下 A 的後面這裡，這兩個好難比。A 跟 D 在比較上，在幻燈片之前的這個部份來講的話，A 比較像是還沒到我們的要求，那 D 有點像畫蛇添足，就是他其實應該已經是到那個要求了，但是他可能就是因為你到了要求你才有能力做 over translation，去犯 over translation 的這個問題，的這個錯。所以在幻燈片之前他們給我的感覺是 D 做得比較好，因為他已經合格，就像我們講那個一關一關一關，他已經到了，他比 A 好可能一關，多走了一關。但是如果從聽者來講的話，就是你不應該犯 over translation 的問題，那他會犯了這個錯。可是如果就口譯來說，他的技巧顯然是比較好。那到幻燈片之後呢，兩個人都差不多，就是在數據上都對，但是都掉了一半。一個就是那個地毯的，地板的這個市場，兩個都沒有講到嘛，那數據是都講了，那 90%的這個也講了，但是兩個人都沒有說出座墊的那個部份。那在數字上 11 億美金這裡，A 是講對了，他講 12 億，但是意思是完全錯誤的。就是我們已經有這樣的營收，但他說是投資，那這邊是完全講錯。幣值也錯，然後你就看到上面的那個嘛。所以如果從後面這個來看的話是兩個一樣糟。所以我想如果真的要比較，以過關度來看的話，D 是比較好的。*（就是掌握的口譯技巧比較多一點？）*對，掌握的比較多。(20:59)

*(overall ranking)*我先看我在筆記上做的東西，因為筆記上寫的是缺點跟優點。那整個這樣看下來的話，第一名我會選擇的是 C，因為他的筆記上我並沒有寫下任何的缺失，我只寫下一個字，memorizing 問號。是因為我覺得 C 做得不錯，讓我覺得他是已經把整個稿子記起來了嗎？所以 C 是第一名。那接下來第二名會是，應該是 B，因為 B 掌握到的訊息很明顯地會比 A 跟 D 來得多，會比其他三個來得多。所以第二個會是 B。然後接下來就是 A D E…在我的觀念裡面我覺得 A 大概會是最後一個。A 是最後一個。所以就只剩下 D 跟 E。A 是最後一個，因為 A 我覺得他的那個口譯技巧是最生疏的。就整個比較起來的話，A 是讓我覺得他還沒有到，他跟其他的四個的 level 可能差一級，所以 A 一定會是最後一個。那我真的要比的話只有 D 跟 E。D 跟 E 的話我會覺得 E 比 D 好，因為兩個人的缺點都差不多一樣，但是 D 有 over translation 的問題，就是他們兩個是同級的，可是 E 沒有犯 over translation 的問題。所以我覺得 E。所以真的排順序起來的話，從第一名開始是 C、B、E、D、A。(23:12)

Appendix I: Sample coding sheet of examiner comments

PCC1	1	delivery / presentation				
			accent			
			coherent	transitions		
			comfortable to listen to	control of breathing	tones	pleasant
			concise			
			convincing, not hesitant	with deliberate steps		
			disconnected sentences			
			focused / unfocused			
			idiomatic usage	word choice / order	regional usage	diction
			pace / rhythm / tempo / fluency	slow / fast	fluent	mumbling pauses
			rendition	polished	literal rendition	simple expression is better
			voice quality / bored	nervousness / tense / steady	sound tired	high pitch
			wordy	fillers / verbal tics		
	2	audience point of view				
			breathing control	sigh (turn off mic)		
			can get more messages			
			concise			
			consistent and coherent speech	transition	confuse audience	
			delivery impact			
			expect to be like a professional	beginner/in-training		
			fragmented			
			idiomatice usage	regional usage to target audience		diction
			more comfortable to listen			
			pace / fast = nervous / slow = relax	willing to keep listening (meaning doesn't really matter)		
			pace / rhythm / tempo / fluency	slow / fast	fluent	mumbling pauses
			polished = easy to understand			
			unfocused			
			unprofessional / girlish fillers			
			warm-up fast	gain audience's confidence		
	2.1	client point of view	faithfulness to SL	company product	purpose of speech	
	3	Fidelity&Completeness				
			consistency	stable/steady		
			content accuracy			
			context			
			different levels of fidelity	more information	significant point / gist	numbers important for business interpreting
			fabrication	over-translation		
			fragment	incomplete stentences		
			keep to the theme	intention of the speaker	introduce company/product	greetings
			make sense	focused / unfocused		
			message weightings			
			mistranslation	meaning errors	correction (strategy)	
			numbers/figures	context of numbers	inconsistent context	
			omission	lagging	summarise	
			slip of tongue			
			terms			
	4	Comprehension&listening				
		foundation abilities	backtrack repair messages	incoherent		
			completely lost			
			comprehension = accuracy			
			comprehension is better			
			did not grasp	fabrication		
			have severe comprehension problem			
			long lags = did not hear			
			misunderstood messages			
			not comprehension problem, but misuse of terms			
			same level of comprehension			
			show consistent comprehension			
			slow pace = poor comprehension	fragmented	incomplete sentences	
	5	Strategy	use of visual aid			
			multi-tasking	know how to distribut efforts		
			paraphrasing	not influenced by source text	use simplier expressions	
			short-term memory			
			EVS range / lagging	speed / pace control		
			approximation of numbers			
			when encounter difficulties and ommissions			
	6	Examiner behaviours				
			attention	memory lag	script use: listen and read	
			bias	accent	know students	personal preferences
			judgement pattern	weightings of criteria	quick / slow decisions	
			marking strategy	look for potentiality		
			past experiences as the audience			
			speculation / EVS / lags / tired	comprehension	strategies / preparation	suspect feeding interpreters misinformation
			use of assessment tool	notes with/without scripts	review recordings	
			FCD approach	fidelity	completeness	delivery
			reverse decision			
			difficult to decide			

References

AIIC. (2000). AIIC Professional Standards. In AIIC (Ed.).

AIIC. (2006). Advice to Students wishing to become Conference Interpreters. Retrieved 17 May, 2009, from http://www.aiic.net/ViewPage.cfm/page56.htm

AIIC. (2009). Code of Professional Ethics. In AIIC (Ed.).

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

Angelelli, C. V. (2006). Validating professional standards and codes - challenges and opportunities. *Interpreting, 8*(2), 175–193.

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

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

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bailey, K. M. (1998). *Learning About Language Assessment : Dilemmas, Decisions, and Directions*: Heinle & Heinle Publishers.

Breeze, R. (2004). Book review: Glenn Fulcher (2003), Testing Second Language Speaking. *TESL-EJ, The Electronic Journal for English as a Second Language, 8*(1), 1-2.

Broadfoot, P. (1996). Assessment and Learning: Power or Partnership? In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, Developments and Statistical Issues*: John Wiley.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1-15.

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

CERI. (2005). *Formative Assessment - improving learning in secondary classrooms*. Paris: OECD Publishing.

Charmaz, K. (2006). *Constructing Grounded Theory*. London: SAGE.

Chernov, G. V. (1979). Semantic aspects of psycholinguistic research in simultaneous interpretation. *Language and Speech, 22*, 277-295.

CILT. (2006). National Occupational Standards in Interpreting: CILT, the National Centre for Languages.

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment: Cambridge: CUP.

Cronback, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 3-17). Hillsdale: Lawrence Erlbaum.

Crooks, T. J. (1988). The impact of classroom evaluation on students. *Review of Educational Research, 5*(4), 438-481.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement* (5 ed.). Englewood Cliffs, NJ: Prentice-Hall.

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.

Ericsson, K. A., & Simon, H. A. (1980). Verbal Reports as Data. *Psychological Review, 87*(3), 215-251.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data*. Cambridge: MIT Press.

Farthing, G. W. (1992). Introspection I: Methods and limitations. In *The psychology of consciousness* (pp. 45-63). Englewood Cliffs, NJ: Prentice Hall.

Frederiksen, J., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

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

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference.* (*11.0 update* (4th ed.) ed.). Boston: Allyn & Bacon.

Gibbs, G., & Simpson, C. (2004). Conditions Under Which Assessment Supports Students Learning. *Learning and Teaching in Higher Education*(1), 3-31.

Gile, D. (1995a). *Basic Concepts and Models for Interpreter and translator Training*. Amsterdam & Philadelphia: John Benjamins.

Gile, D. (1995b). Fidelity Assessment in Consecutive Interpretation: An Experiment. *Target, 7*(1), 151-164.

Gipps, C. (1994). *Beyond Testing: Towards a Theory of Educational Assessment* (2003 Taylor & Francis e-Library ed.). London: Falmer Press.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for Qualitative Research*. Chicago: Aldine.

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

Hartley, A., Mason, I., Peng, G., & Perez, I. (2004). *Peer and self-assessment in conference interpreter training* (project report). York, UK: Subject Centre for Languages, Linguistics and Area Studies, The Higher Education Academy.

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

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

IoL. (1994). *Diploma in Public Service Interpreting Handbook* (2007 ed.). London: Chartered Institute of Linguists (IoL) Educational Trust.

Ivanova, A. (1996). The Use of Retrospection in Research on Simultaneous Interpreting. In S. Tirkkonen-Condit & R. Jääskeläinen (Eds.), *Tapping and Mapping the Processes of Translation and Interpreting* (pp. 27-52). Amsterdam: John Benjamins.

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

Kerlinger, F. N. (1973). *Foundations of Behavioral Research*. New York: Holt, Rinehart and Winston.

Kohn, K., & Kalina, S. (1996). The Strategic Dimension of Interpreting. *META 41*(1), 118-138.

Kopczyński, A. (1994). Quality in Conference Interpreting: Some Progmatic 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. (2001). Conference Interpreting: Quality in the Ears of the User. *Meta, XLVI*, 394-409.

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

Lee, T. H. (2002). Ear Voice Span in English into Korean Simultaneous Interpretation. *Meta, XLVII*(4), 596-606.

Levin, L., C. Langley, A. Lavie, D. Gates, D. Wallace, K. Peterson. (2003). *Domain Specific Speech Acts for Spoken Language Translation.* Paper presented at the 4th SIGdial Workshop on Discourse and Dialogue, Sapporo, Japan.

Li, Z., Cheng, K.-E., Wang, Y., Hiltz, S. R., & Turoff, M. (2001). *Thurstone's Law of Comparative Judgment for Group Support.* Paper presented at the Seventh Americas Conference on Information Systems, Boston, MA.

Liu, M. (2001). Expertise in Simultaneous Interpreting: A Working Memory Analysis, *unpublished PhD dissertation, Graduate School of the University of Texas at Austin*. Austin, Texas: University of Texas.

Liu, M., Chang, C., & Wu, S. (2008). 口譯訓練學校之評估作法: 臺灣與中英美十一校之比較(Interpretation Evaluation Practices: Comparison of Eleven Schools in Taiwan, China, Britain, and the USA). *編譯論叢(Compilation and Translation Review), 1*(1), 1-42.

Luchins, A. S. (1957). Primacy-recency in impression formation In C. I. Hovland (Ed.), *The Order of Presentation* (pp. 33 - 61). New Haven: Yale University Press.

Lumley, T., & 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.

Marzocchi, C. (2005). On norms and ethics in the discourse on interpreting. *The Interpreters' Newsletter, 13*, 87-107.

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, Washingtion: Macmillan.

Miller, C. M. I., & Parlett, M. (1974). *Up to the mark: A study of the examination game*. Guildford: Society for Research into Higher Education.

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 interpreting* (pp. 57-67). Amsterdam & Philadelphia: John Benjamins.

Moser, P. (1995). *Survey on Expectations of Users of Conference Interpretation, translated by J. Mockintosh and C. Stenzl*. Vienna, Austria: AIIC.

Moss, P. A. (1992). Shifting conceptions of validity in educational measuremet: Implications for perfromance assessment. *Review of Educational Research, 62*(3), 229-258.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217-263.

Peng, K.-C. (2006). The Development of Coherence and Quality of Performance in Conference Interpreter Training, *unpublished PhD thesis*. Leeds, England: Centre for Translation Studies, School of Modern Languages and Cultures, University of Leeds.

Phelps, L., Schmitz, C. D., & Boatright, B. (1986). The Effects of Halo and Leniency on Cooperating Teacher Reports Using Likert-Type Rating Scales *Journal of Educational Research, 79*(3), 151-154.

Pöchhacker, F. (1994). Quality Assurance in Simultaneous Interpreting. In C. Dollerup & A. Lindegaard (Eds.), *Teaching Translation and Interpreting 2* (pp. 233-242). Amsterdam & Philadelphia: John Benjamins.

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

Pöchhacker, F. (2004). *Introducing Interpreting Studies*. London & New York: Routledge.

Pöchhacker, F., & Shlesinger, M. (Eds.). (2002). *The Interpreting Studies Reader*: Routledge.

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

Riccardi, A. (2002). Evaluation in interpreting: Macrocriteria and microcriteria. In E. Hung (Ed.), *Teaching Translation and Interpreting 4* (pp. 115-126). Amsterdam & Philadelphia: John Benjamins.

Rowntree, D. (1987). *Assessing Students — how shall we know them?* . London: Kogan Page.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.

Salvia, J., & Ysseldyke, J. E. (1995). *Assessment* (6 ed.). Boston, MA: Houghton Mifflin Company.

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

Schjoldager, A. (1995). Assessment of Simultaneous Interpreting. In *Teaching Translation and Interpreting 3* (pp. 187-195). Amsterdam & Philadelphia: John Benjamins.

Scriven, M. (1967). The methodology of evaluation. In R. E. Stake (Ed.), *Perspectives of curriculum evaluation* (Vol. 1, pp. 39-55). Chicago: Rand McNally.

Shavelson, R. J. (2004). Generalizability Theory. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement*: Academic Press

Shlesinger, M. (1997). Quality in Simultaneous Interpreting. In Y. Gambier, D. Gile & C. Taylor (Eds.), *Conference Interpreting: Current Trends in Research* (pp. 123-131). Amsterdam & Philadelphia: John Benjamins.

Snyder, B. R. (1971). *The Hidden Curriculum*. Cambridge, MA: MIT Press.

StatSoft. (2006). Electronic Statistics Textbook: Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/stathome.html.

Steiner, D. D., & Rain, J. S. (1989). Immediate and delayed primacy and recency effects in performance evaluation. *Journal of Applied Psychology, 74*(1), 136-142.

Stobart, G., & Gipps, C. (1997). *Assessment - A teacher's guide to the issues* (3rd ed.). London: Hodder & Stoughton.

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

Upshur, J. A., & 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.

Vermeiren, H. (2010). The Final Evaluation of Interpreter Performances: A Social Practice. In V. Pellatt, K. Griffiths & S. Wu (Eds.), *Teaching and Testing Interpreting and Interpreting* (pp. 285-300). Bern: Peter Lang.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing 15*(2), 263–287.

Whittington, M. S., López, J., Schley, E., & Fisher, K. (2000). *Using Think-Aloud Protocols to Compare Cognitive Levels of Students and Professors in College Classrooms* Paper presented at the 27th National Agricultural Education Research Conference, San Diego, California.

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

Yang, C. (2000). *口譯教學研究: 理論與實踐 (Reserach on Interpreting Teaching: Theory and Practice)*. Taipei: Fu Jen Catholic University Publishing.