口譯考試之評分行為模式

吳紹銓

摘要

本文以一同步口譯考試之整體研究為本 (Wu, 2013)，報告主題為探索並瞭解口譯考試中的評分行為模式。經研究分析，本文整理出考官的各種評分行為模式，例如，使用考題講稿與記筆記等外顯可觀察的評分行為，還有內在的種種思考模式，包括注意力分配、喜好偏向、職業習慣等，諸如此類之種種評分行為模式皆影響了本研究中三十位考官評分結果的一致性與差異。本研究之成果希望有助於瞭解我們如何評量口譯考生，並進行更多研究，以完善口譯考試設計。

關鍵詞：同步口譯，口譯考試，評分行為模式

吳紹銓，英國新堡大學現代語言學院副教授。

本文部分初稿內容曾發表於「2012 臺灣翻譯研討會－翻譯專業發展與品質提升」，作者在此向該場研討會主持人李憲榮教授、與談人張嘉倩教授及所有與會人士，以及本文的兩位匿名審查者所給予的修正意見致謝。

Examiners' Assessment Behaviours In The Interpreting Examinations

Fred S. Wu

Abstract

Based on a larger study on issues of assessing simultaneous interpreting (Wu, 2013), the aim of this paper focuses on exploring and understanding the examiners' assessment behaviours. A range of assessment behaviours was identified among the thirty examiners in this study. In addition to the more observable external behaviours, such as the use of speech script and note-taking as assessment tools, this study also identified some internal behaviour, such as examiners' attention, bias, and professional habits, which may explain the consistent and inconsistent test results of the thirty examiners. These findings may hopefully lead to a better understanding on how we assess student interpreters for more future research, and contribute to a better test design of the interpreting examinations.

Key words: simultaneous interpreting, interpreting examination, assessment behaviour

---

Fred S. Wu, Senior Lecturer, School of Modern Languages, Newcastle University, United Kingdom

1. Introduction

The evaluation of the appropriateness of a test and its application requires the judgement of professionals for validity reasons. 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. However, when making a judgment in a performance assessment such as interpreting, it is in nature a subjective one, and subjective judgment has long been criticized as less reliable (Campbell and Hale, 2003: 212). Professional judgement alone, therefore, may not be a sufficient basis for decision-making in the examinations. 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 (ibid: 104).

The test design and assessment criteria for evaluating interpreters' performance, nevertheless, have been considered as "intuitive" (ibid: 211); the way interpreter examiners apply assessment criteria has also been described as "fuzzy" (Sawyer, 2004: 185). The test items and test procedures in many interpreter educational institutions have been often designed and administered with little or no basis of empirical studies (Liu, M., Chang, C. and Wu, S., 2008: 35). Some institutions in Liu et al.'s survey study 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. 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, and so on. There were common grounds and there have been efforts to improve the assessment methods. However, the guidelines were often found to be difficult to follow because of the need to retain the authenticity of the task in the performance-based assessment, especially in live panel examinations. The methods and instruments for interpreting assessment varied among different interpreter teachers, examination boards and training institutions, and the approaches have often been based on subjective judgement rather than on empirical data (ibid: 17-18, 34-35).

Therefore, the concern about how consistent the examiners in the interpreting examinations judge interpreting performances, especially within the educational

context, still remains. It is necessary to look into these concerns in the interpreting examiners with more systematic studies.

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). For example, 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. 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 and Hale, 2003: 221).

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 and McNamara, 1993: 3). A perfect examiner that is always consistent and reliable is almost impossible to find, and it is the ordinary examiners that may present problems in a test, such as halo, overall severity/leniency, central tendency, and random errors in their judgement (ibid). These problems, or rater effect, will have an influence on the results of many performance-based assessments, making the assessment procedure become less reliable and threatening the validity of the test (Eckes, 2005: 197).

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, Lynch, and Mason, 1995; Eckes, 2005; Fulcher, 2003; Lumley and McNamara, 1993; Upshur and 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, Clapham, and Wall, 1995: 105).

Since subjective judgement is at the core of current assessment practice of the interpreting examinations, a logical step for study is to explore and understand the examiners' assessment behaviours. Taking the background and rationale above, a research study was conducted to explore and understand how individual examiners perceive the interpreting performances in a simultaneous interpreting examination, such as the use of assessment criteria and how they make the judgments. The overall study method and main findings are published in a book chapter (Wu, 2013), including a summarisation of the assessment behaviours of the examiners in the interpreting examinations as part of the main study findings of the research study (ibid: 26-28). Based on the findings, this paper aims to discuss in more details the examiners' assessment behaviours and expand the basic conceptual model of interpreting examinations (ibid: 29) in order to have a better understanding of how we assess students in the interpreting examinations.

2. Study method and results

For data collection, the study conducted a simulated exam of simultaneous interpreting and invited thirty examiners to judge five postgraduate student interpreters' performances from video recordings, and recorded the examiners verbal comments during and immediately after they made a judgement on the students' interpreting performances. The study method and procedures were described in Wu (2013: 17-20) as part of the main study. This paper summarises here the participant examiners' background below, and explains the study method for analysing the examiners' behaviours.

There are in total thirty examiners who participated in this study. For contrastive analysis, the thirty examiners came from three professional backgrounds as below.

- 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

They included 19 interpreter examiners and 11 non-interpreter examiners. All of their working languages are Mandarin Chinese and English, with Mandarin being the first language of all the examiners except one who was based in the UK. The examiners were asked to listen to five examination recordings of student interpreters, which were selected from an exam archive. The selection was based on a mark range from 50s (pass) to 70s (distinction) in the hope to illicit a wider range of responses of judgments from the participant examiners. The five students were coded from A to E.

Thurstone's Method of Paired Comparison (Thurstone, 1959) was adopted to monitor the consistency level of the examiners' judgment results; the examiners were asked to compare the student interpreters in pairs, and to think aloud their judgment processes, and interviewed when deciding which performance was better. Then, the examiners' verbal comments were recorded, transcribed and coded for analysis in order to extract any salient assessment behaviours during the judgment process (see Wu, 2013: 24).

Some behaviours are easy to observe, and can be referred to as the examiner's external behaviour, for example using the assessment tools. However, 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 analyse human thoughts, is verbal report analysis (Kucan and Beck, 1997 in Whittington, López, Schley and 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 and 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 articulate their thinking, i.e. think aloud, that leads to the solution of a problem at hand. Analysing such verbal reports may help researchers to understand how people think in relation to the task that they do.

Some scholars (Conrad, F., Blair J., and Tracy E., 1999) also discussed the possibility that the act of verbal report may alter the thinking being reported, which may in turn lead to degrading or distorting the main task being performed. They presented a counter-argument that 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 (1980, 1993)

tested the validity of this argument and they found that the act of introspection did not affect their test subjects' mental processes: subjects went through the same steps whether they concurrently described what they were doing, retrospectively described it, or did neither. This test result suggested that introspection can be practiced in reliable ways as a research method (ibid in Conrad et. al., 1999).

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.

Table 1 Types of Examiner Behaviour

Types	Conceptual properties
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, I didn't write it down, review recordings, noting errors on the script, judging from notes, let me compare them 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, 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

(continued)

Table 1 (continued)

Internal **Behaviour**	interpreter's background – word choice **professionally-referenced standards** guessing comprehension, 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

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.

Based on the coding principle of the Grounded Theory (Bryman 2004: 401-408), therefore, when a distinctive idea or concept was identified in the examiners' comments in this study, 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 judgements. In achieving the study aim, the coding process focused on any conceptual key words from which inferences can be drawn on how the examiner judged the interpreting performances. After the line-by-line coding of the thirty examiners' comments, the coded concepts were then compared and collated with one another; similar concepts were grouped into categories. Table 1 presents how the conceptual properties 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 (see Table 1 and Wu, 2013: 23-26). Through verbal report analysis and the coding process in this study, therefore, the examiners' interview comments and the concepts extracted from the interview data may provide a window to explore and understand the examiners' internal behaviours as well as various factors that may affect them.

The assessment behaviours are closely linked to the use of criteria for judgement in the interpreting examinations. Wu (2013) proposed a basic conceptual model of interpreting examinations (the IE model) to illustrate two dimensions in the interpreting examinations: assessment criteria dimension and assessment behaviour dimension. The examiners' behaviours are illustrated as the Speaker-Examiner-Audience triangle in the lower part of the model as shown in Figure 1.

Based on the identified conceptual properties as shown in Table 1, the behaviour triangle of the IE model (Figure 1) will be further discussed and expanded in the hope to better understand the examiners' assessment behaviours in the interpreting examinations.

Figure 2 shows the revised behaviour triangle of the IE model, which illustrates in details how the identified assessment behaviours relate to and interact with each other. The judgement 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, and two professionally-referenced behaviours – the *condensation norm* (i.e. interpreter's reduction strategy), and *situational weighting*

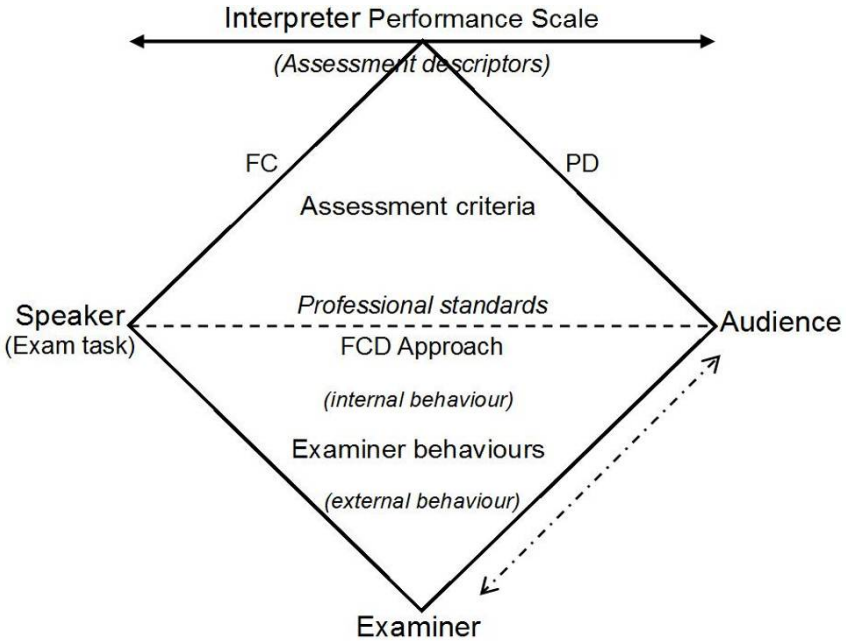

Figure 1 The basic conceptual model of interpreting examinations (Wu, 2013: 29)

FC: Fidelity and Completeness, PD: Presentation and Delivery

FCD Approach: Fidelity-Completeness-Delivery Approach

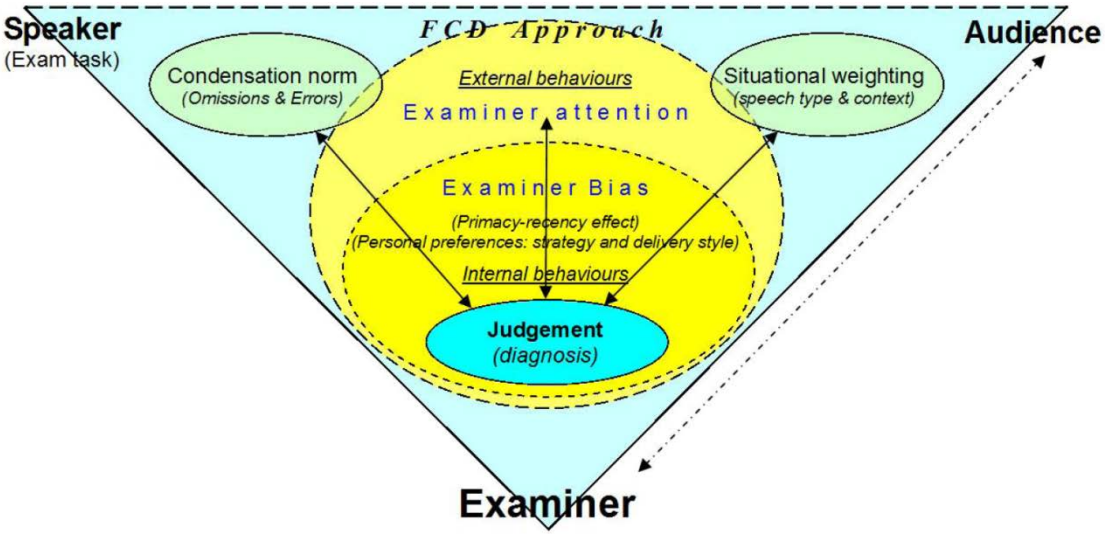

Figure 2 Revised behaviour dimension of the IE model

of the source speech type and context. 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* and *examiner bias*, which includes the *primacy-recency effect* and *personal preferences* (see Table 1). In addition, the examiners may assess the students from the audience point of view, thus, playing a dual role in an interpreting examination (Wu, 2013: 26-28).

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 2 for illustration.

3 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.

3.1 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 2, which is close to the Professional standard in the

criteria dimension as illustrated in Figure 1. They may be considered as the examiners' collective assessment behaviours in the interpreting examinations.

Fidelity and Completeness (FC) and Presentation and Delivery (PD) are the two primary assessment criteria that the examiners used when assessing student interpreters, which are illustrated on the two slops of the criteria triangle as shown in Figure 1. 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. FC is also weighted more than PD when both criteria are considered for making a judgement. Comments 1 and 2 are two examples of the FCD judgement approach.

> **Comment 1 (translation)[1]:** […] I will pay more attention to see if there are 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.

> **Comment 2 (translation)[2]:** 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.

---

[1] **Comment 1** 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。

[2] **Comment 2** 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。

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 2. 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.

Juggling judgements between omissions and errors in student interpreters' performances is common for examiners when they are applying the Fidelity and Completeness criterion. The findings of this study also show that the examiners would 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. 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).

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 3 and 4 below illustrate how the examiners in this study applied this condensation norm when assessing student interpreters.

> **Comment 3 (translation)[3]:** D is worse than C. First, she (D) is not fluent enough; second, she omitted more messages, […]. Compared with C, however,

[3] **Comment 3** 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。

because she (D) omitted a lot, there seems to be less error (in D's delivery).

The underlined part of Comment 3 illustrates the examiner's view of the relationship between omissions and errors when interpreting. In Comment 4, when comparing Students D and E, the same examiner further elaborated on which is the more serious – omission or error. The examiner concluded that overall Student D was better.

> **Comment 4 (translation)[4]:** It's because that although she (D) 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 5 sums up this assessment approach in weighing omissions against errors.

> **Comment 5 (translation)[5]:** 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, 1995: 200). In this study, some examiners also made a distinction between not understanding the message and not hearing the message at all. 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.

According to the examiners' comments, 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 validation (Alderson et al., 1995).

---

[4] **Comment 4** in source text Chinese: 因為他(D)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。

[5] **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。

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, the background of the interpreter examiners varies and inconsistencies in their judgements are bound to happen, which is a cause for concern (Sawyer, 2004: 184). For example, this study found that 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.

Nevertheless, some non-interpreter examiners also share similar judgement patterns to the interpreter examiners. Therefore, it appears that more factors than just the examiners' professional background will maintain consistency or 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 2. These two types of assessment behaviours are more personally related to the examiners, which will be discussed in the sections below.

3.2 Examiner bias

As shown in Figure 2, the inner circle of examiner bias includes two biases identified: the *primacy-recency effect* and *personal preferences*. The examiners will be influenced by these biases, consciously or unconsciously, when judging or diagnosing student interpreters' performances.

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). Theoretically speaking, therefore, this primacy-recency effect is likely to happen to most, if not all, examiners. In this study, those examiners who reviewed the recordings are more likely to notice this effect. After examining the five student interpreters, they 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 6 illustrates a typical comment relating to such a view.

> **Comment 6 (translation)**[6]**:** In fact when I listened for the second time, I had some doubts about my previous judgments. The notes that I had made previously were more of a general impression, which I feel had some "anaesthetic" effect. The students did not interpret particularly well, and the sequence of listening to them made some difference [in judgments]. 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.

Since the primacy-recency effect is a cognitive bias, it 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 and 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. 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

[6] **Comment 6** 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。

student interpreters from a dual role perspective as Examiner and Audience as indicated as the double-arrowed dotted line in Figure 2.

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. 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 self-contradicted decisions during the judgement process, i.e. intra-rater inconsistency. Comments 7-9 below illustrate some typical examples to show how an examiner's preference influences the decisions made.

> **Comment 7 (translation)**[7]**:** Overall speaking, both (students) had a lot of mistakes, but I like E's interpretation better because I feel that E was more fluent, not in such a hurry. […] In this sense, therefore, I think Eileen is the better one.

> **Comment 8 (translation)**[8]**:** As for E, I cannot stand listening to her. […] 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.

Both examiners in Comments 7 and 8 obviously made the decision based on the same criteria, i.e. delivery, but one favoured E and the other couldn't stand E's delivery style. It is clear that the examiners had preferences for the interpreter's delivery style, which played a part in making their judgements.

The examiner in Comment 9 also did not like Student E's delivery style. This examiner's personal preference was so strong that it was enough to influence the

[7] **Comment 7** 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。

[8] **Comment 8** in source text Chinese: E 的話我看，我很受不了她說話的樣子，[…]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。

examiner to deviate from the FCD approach when comparing Students D and E. The examiner made the following comment.

> **Comment 9 (translation)**[9]**:** 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.

The comment shows that even though this examiner knew that E "managed to make a lot of points", she still would not pick E because of E's delivery style.

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.

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 (Alderson et at., 1995: 105). However, examiner training can only reduce "extreme differences" in assessment behaviours and the examiner variability cannot be totally eliminated (Lumley and 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 and 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

[9] **Comment 13** 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。

examination procedures that help avoid or minimise the potential harm from the examiner bias.

3.3 Examiner attention

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, 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 assessing simultaneous interpreting, therefore, the examiners may often resort to holistic marking as a result 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 above. Comments 10-12 below are examples of typical comments that show the limited attention span of the examiners, and how they may make a judgement by impression.

> **Comment 10 (translation)**[10]**:** 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.

[10] **Comment 10** 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。

Comment 11 (translation)[11]**:** "I didn't take many notes about C and E. I *felt* that E is better" (*emphasis added*).

Comment 12 (translation)[12]**:** I didn't write them down, but because…sorry, I didn't hear very clearly why (they were wrong) because if she (A) was wrong from the beginning to the end in the process of the examination, I wouldn't bother to remember the details.

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 among the juries that are based only on subjective judgements with no evidential support may often lead to less productive results. 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 the 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 on the scripts 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.

[11] **Comment 11** in source text Chinese: C 跟 E 我沒有記下太多筆記，我*覺得* E 比較好。

[12] **Comment 12** 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。

3.4 External assessment behaviour

The external behaviour mainly concerns the use of assessment tools. As discussed above, using practical assessment instruments like the source speech script for note taking may be a good support to the examiners when working under high cognitive and memory load, such as assessing simultaneous interpreting. 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.

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. 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, the rehearsal practice and the discussion of the suitability of the examination task might only happen shortly before the interpreting 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.

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, and the process should ideally be documented for future reference. By doing so, it leaves more time to improve the examination tasks when necessary. In the meantime, 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. Therefore, documentation of the test design on various considerations, such as those discussed above, will be invaluable over time.

4. Conclusion

Although the results of this study may not be directly generalised to real-life examination panels where a number of examiners are present, the findings of this study give useful pointers in understanding how individual examiners may assess student interpreters (Wu, 2013: 30-31). The revised behaviour dimension of the IE model (Figure 2) may also serve as a conceptual map to help us better understand how the examiners judge and diagnose student interpreters' performances through some external and internal influences, ranging from the use of assessment tools to support the examiners' attention span, to the dynamic interactions between personal biases and professional norms. The dynamics of these influences then become the base to support and balance the criteria dimension in the upper part of the IE model (Figure 1) (Wu, 2013: 29). With this knowledge, it is hoped that an improved test design and

examiner-friendly marking procedures may be developed to help achieve a more reliable result of the interpreting examinations.

Reference

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

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.

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.

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

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

Conrad, F., Blair J., and Tracy E. (1999) "Verbal Reports are Data! A Theoretical Approach to Cognitive Interviews", Proceedings of the Federal Committee on Statistical Methodology Research Conference. Retrieved May 13, 2010, from http://www.bls.gov/osmr/pdf/st990240.pdf

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.

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

Gile, D. (1995). *Basic Concepts and Models for Interpreter and translator Training*. Amsterdam & Philadelphia: John Benjamins.

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

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

Marzocchi, C. (2005). On norms and ethics in the discourse on interpreting. *The Interpreters' Newsletter*, 13, 87-107.

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

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.

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.

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.

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.

Wu, F. S. (2013). How Do We Assess Students in the Interpreting Examinations? In D. Tsagari and R. Deemter (Eds). *Assessment Issues in Language Translation and Interpreting* (pp. 15-33). Bern: Peter Lang.