Han, C. (2017). Using analytic rating scales to assess English–Chinese bi-directional interpreting: A longitudinal Rasch analysis of scale utility and rater behaviour. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *16*, 196–215.

Using analytic rating scales to assess English–Chinese bi-directional interpreting: A longitudinal Rasch analysis of scale utility and rater behaviour[1]

Chao Han

Southwest University
chao.research@gmail.com /chaohan@swu.edu.cn

Descriptor-based analytic rating scales have been used increasingly to assess interpreting quality. However, little empirical evidence is available to support unequivocally the effectiveness of rating scales and rater reliability. This longitudinal study therefore attempts to provide insight into scale utility and rater behaviour in the assessment of English–Chinese interpreting performance, using multifaceted Rasch measurement. Specifically, the study focuses on criterion or scale difficulty, scale effectiveness, rater severity or leniency and rater self-consistency between English–Chinese interpreting and over three time points. The research results are discussed, highlighting the utility of analytic rating scales and the variability of rater behaviour in assessing interpreting. The results also have implications for developing reliable, valid and practical instruments to assess the quality of interpreting.

1. Introduction

Interpreting quality constitutes a crucial part of maintaining professional interpreting standards (Diriker, 2015), a major pedagogical concern (Sawyer, 2004) and a timeless topic for researchers (Pöchhacker, 2001). Needless to say, reliable, valid and practical assessment is also an abiding area of research. One way to assess the quality of interpreting is what measurement specialists call an "atomistic method", one in which propositional and (para)linguistic features such as errors, omissions and pauses are counted and tallied as frequency data. Barik (1971) and Gerver (1969/2002), for instance, are among some of the earliest researchers who applied the atomistic method to evaluating spoken-language renditions. This time-honoured method prevails today in high-stakes settings such as professional certification testing. Australia's National Accreditation Authority for Translators and Interpreters (NAATI), for example, has used error deduction/analysis rating, a specific form of the atomistic method, to assess translation for more than 30 years (Turner, Lai, & Huang, 2010). A similar scoring practice is also adopted in the China Accreditation Test for Translators and Interpreters (CATTI) and the United States Federal Court Interpreter Certification Examination (FCICE).

Undoubtedly, the atomistic method has a number of advantages. Chief among them is that the method helps assessors to gain an insight into the nuances of an instance of interpreting at a local level. Another advantage is its intuitive nature: it seems easy to understand and use. However, scholarly concerns about the legitimacy, replicability and practicality of the atomistic method are mounting. First, the method focuses on lexico-semantic details but does not deal with the more global discoursal and pragmatic aspects of an instance of interpreting, which are arguably equally important (Clifford, 2001). As a result of the emphasis on lexical items, some assessors become fixated on error classification to such an extent that they may not develop a holistic appreciation of individual instances of interpreting (Turner et al., 2010). Secondly, the categorization and definition of errors and omissions seem to vary from one author to another, and are therefore not always replicable (Setton & Motta, 2007). Indeed, substantial inter-assessor variability may exist when it comes to identifying and categorizing errors and omissions (Gile, 1999). Finally, the method is time-consuming and "extremely tedious" (Liu, 2013, p. 173), and so it may not be an optimal choice for large-scale testing enterprises such as NAATI and CATTI.

Alternative approaches to quality assessment have a long history in non-interpreting translation modalities. For example, the use of analytic rating scales can be traced back more than 50

years to John B. Carroll (1966), who trialled them to measure the quality of machine translation (MT). Recently, descriptor-based analytic scales have emerged as a promising alternative to the traditional atomistic method in interpreting (see Angelelli, 2009; Han, 2015a; Lee, 2015; Liu, 2013). Confidence in using rating scales to assess interpreting is beginning to increase, as their potential to contribute to reliable, valid and practical assessment comes to be more fully appreciated. Much of the confidence is, however, derived indirectly from abundant psychometric evidence of the general value of rating scales reported in psychological, educational and language-testing literature (e.g., Xi & Mollaun, 2006). Although encouraging results pointing to the effectiveness of rating scales in interpreting assessment are available, as can be seen in the literature review below, such evidence is far from conclusive. There is a need to provide scaffolding based on rigorous research; this empirical study aims to build such scaffolding to support the utility of descriptor-based scales in the assessment of interpreting.

2. Quality assessment of interpreting: Using analytic rating scales

Using descriptor-based analytic rating scales to assess interpreting quality is gaining currency in interpreter certification testing (Angelelli, 2009; Certification Commission for Healthcare Interpreters (CCHI), 2011; Han, 2015a, 2016a; IoL Educational Trust, 2010; Jacobson, 2009; Liu, 2013; Turner et al., 2010; Wu, Liu, & Liao, 2013), in interpreter educational assessment (Bontempo & Hutchinson, 2011; Lee, 2008; Tiselius, 2009; Wang, 2011; Wang, Napier, Goswell, & Carmichael, 2015; Zhao & Dong, 2013) and in interpreting research (Cheung, 2014; Lin, Chang, & Kuo, 2013; McDermid, 2014). This popular trend, however, does not necessarily confirm the utility and functionality of descriptor-based rating scales. Indeed, a critical review of the interpreting literature suggests that overall the use of rating scales is fraught with uncertainty about their usefulness, despite some positive, yet preliminary and limited, evidence to the contrary.

A primary source of uncertainty derives from the unavailability of rigorous empirical evidence to substantiate the effectiveness of rating scales and, as a result, improved rater reliability. Much of the available literature pertains to theoretically oriented design of rubrics and rating scales without concomitant empirical validation (Angelelli, 2009; Bontempo & Hutchinson, 2011; Clifford, 2001; IoL Educational Trust, 2010; Jacobson, 2009; Wang, 2011). Jacobson (2009), for example, drew upon interactional sociolinguistics and conversation analysis to propose an analytic scoring rubric for assessing the interactional competence of community interpreters in healthcare settings in the United States. However, Jacobson (2009) observed that the rubric needed to be examined regarding whether it was consistent in its results, and was therefore "subject to revision as findings from empirical research come to light" (p. 67). Similarly, Clifford (2001) designed an assessment rubric for assessing interpreting based on discourse theory, but called for "some much needed work on quality assurance in interpreter assessment, by focussing on measures of validity, reliability, equity, utility or comparability" (p. 376). A rare yet commendable endeavour, reported in Lee (2015), concerns a rigorous statistical approach to developing an analytic rating scale in assessing interpreting. However, as he suggested, the utility of the rating scale needs further testing and validation, using different samples of interpreter.

Another body of interpreting literature provides some useful, though indirect, evidence on the utility of descriptor-based rating scales (Cheung, 2014; Lin et al., 2013; McDermid, 2014). These empirical studies represent substantive enquiries into interpreting/translational phenomena, in which interpreting quality is a key variable that is measured by raters using descriptor-based scales. For instance, Cheung (2014) investigated whether using anglicized numerical denominations as a coping tactic could improve the quality of English–Chinese simultaneous interpreting (SI) (i.e., fluency and accuracy). Studies of this kind have produced some indirect evidence regarding scale utility. Lin et al. (2013) and McDermid (2014) reported very high inter-rater reliability estimates (e.g., Pearson's $r > 0.90$), suggesting that raters were consistent in using rating scales to award scores. Such evidence, however, is of limited value, given that it does not directly focus on whether the scales have functioned appropriately.

Moreover, a small collection of literature concerns post-hoc validation of interpreting performance tests in which descriptor-based scales were used or trialled (CCHI, 2011; Liu, 2013; Turner et al., 2010; Wu et al., 2013; Zhao & Dong, 2013). Encouraging preliminary results supported the usefulness of rating scales. For example, in the study by Turner et al. (2010) descriptor-based rating was trialled to examine its effectiveness vis-à-vis an error analysis or deduction method. It was found that overall the graded series of holistic descriptors functioned properly to rank-order test candidates. Converging evidence, reported by Liu (2013), also shows that fidelity scores derived from scale- and proposition-based rating achieved a high correlation (Pearson's $r = 0.945$) at a statistically significant level ($p < 0.001$). Liu (2013) even observed that

> the rather easy-to-do scale-based rating for fidelity can be used as a substitute for the highly rigorous yet extremely tedious proposition-based rating when judging accuracy in interpretation. (p. 173)

Contrary to these results, Gile (1999) found a lack of a straightforward link between the number of errors and omissions identified and the general ratings given by assessors when it comes to assessing the fidelity of English–French SI. Han (2015b) also reported that in evaluating the fluency of English–Chinese SI, paralinguistic indicators such as the number of unfilled pauses, mean length of run and the number of false starts did not correlate strongly with the rater-generated fluency ratings, with the absolute values of Pearson's r ranging from 0.06 to 0.62. Given these contradictory findings, the uncertainty lingers.

Apart from these correlational findings, several studies examined rater behaviour (CCHI, 2011; Wu et al., 2013; Zhao & Dong, 2013). It is noteworthy that Zhao and Dong (2013) and Wu et al. (2013) applied multifaceted Rasch measurement to examine rater behaviour in using rating scales. Zhao and Dong (2013) found that on average it was slightly more difficult for the students to score highly on accuracy criterion or scale than on target language (TL) expression in English–Chinese consecutive interpreting (CI (i.e., bi-directional CI). Wu et al. (2013) also observed increased difficulty for the fidelity criterion or scale than for the delivery criterion or scale in English–Chinese CI.

Perhaps the most promising line of research that has the potential to provide insight into scale utility and rater behaviour pertains to the empirical validation of newly devised rating scales or a series of graded rubrics (Han, 2015a; Lee, 2008; Tiselius, 2009; Wang et al., 2015). On a positive note, raters found the scales easy to use (Tiselius, 2009) and perceived them as a means of enhancing rating consistency (Lee, 2008). However, uncertainty has also emerged from this research: in particular, rater reliability estimates pointed to a possible differential functioning of rating scales for different types of rater. Wang et al. (2015) reported higher inter-rater reliability between two interpreter educators than between each interpreter educator and an interpreting practitioner. Tiselius (2009) also found that, using descriptor-based scales, the PR group (i.e., professional interpreters as raters) achieved higher inter-rater reliability than the SR group (i.e., student interpreters as raters). Interestingly, a conflicting result was observed by Lee (2008) in which higher inter-rater reliability was obtained for the SR raters than for the PR raters.

Although the above studies have contributed to an enhanced understanding of scale utility and rater behaviour, further insights need to be gained to provide clearer answers. The review identifies several aspects of the previous studies that could benefit from further methodological modification. First, despite the fact that multiple criteria or scales are usually used, there has seldom been an explicit attempt to examine the relative difficulty of each criterion or scale (e.g., Zhao & Dong, 2013). More importantly, most of the studies do not directly concern themselves with the effectiveness of rating scales, that is, whether they have exhibited desirable psychometric properties, which should be subjected to empirical analysis.

Secondly, in analysing rater behaviour, a majority of the studies use inter-rater reliability estimates, particularly Pearson's r, which has several limitations (for details, see Han, 2016b). Chief among them is that an inter-rater reliability coefficient between rater A and rater B is not a unique coefficient for either of the raters. Another limitation is that Pearson's r merely indicates whether raters rank-order performances in a similar fashion (see, for example, Multon, 2010). As a result,

even if inter-rater reliability (Pearson's *r*) were high, a statistically significant difference could still exist. In Wang et al.'s (2015) study, for example, although raters Jamie and Morgan achieved a relatively high correlation ($r = 0.87$) between their total scores, ANOVA-based rater severity analysis suggested that on average Morgan was much more severe than Jamie ($p < 0.05$).

Thirdly, not all studies are based on a fully crossed measurement design. In Wang et al. (2015), while two raters (Ashley and Morgan) assessed all the instances of interpreting, the third rater (Jamie) worked only on a subset (about 50%) of them. This incomplete rating design or missing data from Jamie prevents the researchers from mining the data thoroughly: much of the statistical analyses were restricted to the common set of instances of interpreting assessed by all three raters. An associated issue is the use of a single task for each interpreting direction (see Tiselius, 2009; Wang et al., 2015; Zhao & Dong, 2013). Moreover, in some of the studies, only unidirectional interpreting is required (Han, 2015a; Lee, 2008). The mono-task design and the uni-directionality do not allow for observing and comparing how scales would function and how raters would behave in different tasks and between a given language pair.

Finally, the previous studies are exclusively based on one-off assessment. Such a design offers a snapshot of scale utility and rater behaviour at a specific time point, but is unable to capture developmental trends or to generate more stable results. Further research is therefore needed before meaningful conclusions about the reliability of scale-based assessment can be drawn.

3. Research questions

Against this backdrop, this longitudinal study draws upon a fully crossed measurement design to explore and track, based on multifaceted Rasch measurement, how descriptor-based rating scales function and how raters behave in assessing bi-directional interpreting over time. Specifically, three major research questions (RQ) were investigated:

RQ1**:** What was the level of difficulty of the three criteria or scales used in the present study: information completeness (InfoCom), fluency of delivery (FluDel), and TL quality (TLQual)?
RQ2: Did the three descriptor-based rating scales (i.e., InfoCom, FluDel and TLQual) function appropriately?
RQ3: How did the raters in the study behave in assessing interpreting? Specifically, (a) what was the rater severity or leniency? (b) were the raters self-consistent in using the scales?

It needs to be pointed out that all the RQs were examined longitudinally and in relation to both interpreting directions (English–Chinese and vice versa).

4. Background to the study

The study relates to an obligatory third-year undergraduate course in English–Chinese CI. A total of 38 students from two classes registered for the course; they were all enrolled in a Bachelor of Arts programme, majoring in English–Chinese translation. Average age 21 years, a majority of the students were female ($n = 32$) and the remainder were male ($n = 6$). All the students were native Chinese who were learning English as a foreign language. The course was run over a period of ten contact weeks in which the students and the teacher (i.e., the author) met once a week for one-and-a-half hours.

The students participated in three performance assessments in the fourth, the ninth and the tenth week. Each assessment comprised six tasks: three tasks for each direction (C–E and E–C). Three days before each assessment, the students were briefed about topics and themes to be interpreted and also given a source-language (SL) word list for preparation (e.g., finding TL equivalents). Overall, each task featured a two-minute generalist and moderately paced speech (approximately 110 words per minute). Each speech was divided into smaller segments of approximately 40 seconds. The students were allowed to take notes and interpreted segment by

segment. All the CI performances were audio-recorded. A total of 228 recordings (i.e., 38 students × 6 tasks) were therefore generated on each occasion.

5. Methods

5.1. Raters

A panel of six raters was recruited, including two interpreting teachers and four teaching assistants. The teachers, both in their early thirties, were interpreters by training and were university lecturers in English–Chinese interpreting (i.e., Raters 05 and 06). The four teaching assistants, averaging 23 years of age, were postgraduate students pursuing a master's degree in Translation and Interpreting (i.e., Raters 01, 02, 03 and 04). In addition, all the raters had Mandarin Chinese as their native language and English as a second language.

5.2. Descriptor-based rating scales

Three rating scales were used to assess three aspects of interpreting: information completeness (InfoCom), that is, to what extent source-text propositional content is interpreted, fluency of delivery (FluDel), that is, to what extent disfluencies such as un/filled pauses, long silences and fillers are present in TL renditions, and TLQual, that is, to what extent TL expressions are idiomatic and grammatically correct. The descriptors in the scales were adapted from Han (2015a, 2016a), but were also revised based on the students' feedback and suggestions. In addition, each of the scales is flexible in the sense that it can be simplified to a four-band scale by collapsing two neighbouring points into a single score band.[2]

5.3. Procedure

Before formal rating, a rater training session was held that lasted approximately three hours. In the training, the author first gave each rater a copy of the scales so that they could familiarize themselves with scale structure and descriptors on their own. Next, the author explained key terms in the descriptors such as deviation, omission and disfluencies, and provided illustrative examples to the raters. In particular, the raters were asked to heed the relevant features of a given instance of interpreting, identify evidence to match scale descriptors and categorize the interpreting into one of four scale bands before awarding a band score. When awarding scores, the raters were also asked to proceed from InfoCom to FluDel and TLQual, a rating strategy identified by Wu (2010) based on a sample of 30 raters. Meanwhile, the raters were encouraged to ask any questions and air their views on the scales, so that potential misunderstanding and confusion could be addressed. Then, all the raters were provided with 15 randomly selected recordings of interpreting for trial rating. After each round of rating, the raters compared the scores with one another. When disagreement arose, each rater explained why they had scored in a particular way. By doing so, it was believed they could become aware of potential differences, gain confidence in using the scales and try to reconcile their rating with that of other raters. It needs to be pointed out that the rater training was provided only once before the first formal rating in Week 4. Providing such intensive training and orientation is generally desirable (see Wang et al. 2015) because the raters recruited in the study were not familiar with the scale-based assessment of interpreting, particularly with the scalar dimensions, descriptors and structure. A lack of basic training could also lead to the validity of the assessment outcomes being undermined.

During the formal rating, all the raters gathered together but worked independently. Each rater assessed a batch of 20–25 randomly selected recordings before taking a 15-minute break. They were also provided with SL texts to help them check the renditions against the original content. This

rating arrangement was operationalized in Week 4 and repeated in Weeks 9 and 10; each rating session lasted two consecutive days.

5.4. Measurement design

A fully crossed measurement design was operationalized in which each rater assessed all the instances of interpreting in each assessment (i.e., Weeks 4, 9 and 10), using the three rating scales. This is regarded as an optimum design, as the data points are completely connected (Schumacker, 1999). Consequently, on each occasion, a total of 2,052 data points were generated for each direction (i.e., 38 students × 3 tasks × 3 criteria × 6 raters).

5.5. Data analysis

Multifaceted Rasch measurement (MFRM) was implemented to mine the data. The study highlights four assessment facets: (a) students, (b) raters, (c) tasks and (d) criteria or scales. Through joint maximum likelihood estimation procedures, the MFRM generates calibrated estimates for all the elements in each assessment facet in a common equal-interval metric, known as log odd units or logits. This results in the creation of a single frame of reference for analysing results (for details, see Linacre, 2013). Han (2015a) and Han and Slatyer (2016) have called for the use of MFRM to analyse the scale-based, rater-mediated assessment of interpreting.

A Rasch model variant called the "Partial Credit Model" (PCM) was used (Masters, 1982). The PCM assumes that each rating scale had its own distinctive structure for each assessment criterion. Each of the fully crossed data sets was subjected to separate MFRM analysis, based on the FACETS 3.71.0 program (Linacre, 2013), before being collated to gain a longitudinal understanding of scale utility and rater behaviour. To ensure the clarity and reader-friendliness of the analysis, only those Rasch-calibrated measures and statistics that help answer RQs are presented.

6. Results

6.1. Difficulty of criterion or scale

The MFRM analysis estimated criterion or scale difficulty in logit (not based on raw scores). Table 1 summarizes logit estimates of difficulty of each criterion or scale, for each direction and each time point. A larger logit value indicates an increasing level of difficulty for a given criterion or scale; a logit value of 0.0 means that the level of difficulty of a criterion or scale is at an average level.

Table 1: Logit estimates of level of difficulty of criterion or scale

Directionality	Criteria	Difficulty estimate (in logit)		
		Week 4	Week 9	Week 10
English–Chinese	InfoCom	0.11	0.29	0.10
	FluDel	0.01	−0.01	0.02
	TLQual	−0.12	−0.28	−0.12
Chinese–English	InfoCom	−0.62	−0.69	−0.72
	FluDel	0.36	0.43	0.38
	TLQual	0.26	0.25	0.34

To gain insight into the data, the logit estimates were plotted visually. As can be seen in Figure 1a, for English–Chinese CI, in Week 4 the InfoCom criterion or scale was the most difficult (logit = 0.11), whereas TLQual was the easiest (logit = −0.12). In other words, with all other assessment facets remaining constant, the students scored lowest on InfoCom but highest on TLQual. Notationally, regarding the level of difficulty, it could be described as "InfoCom > FluDel > TLQual". This pattern was also consistent over time, from Week 4 to 10. As reflected in Table 1, the logit differences between the three criteria/scales were fairly small, ranging from 0.22 to 0.57.

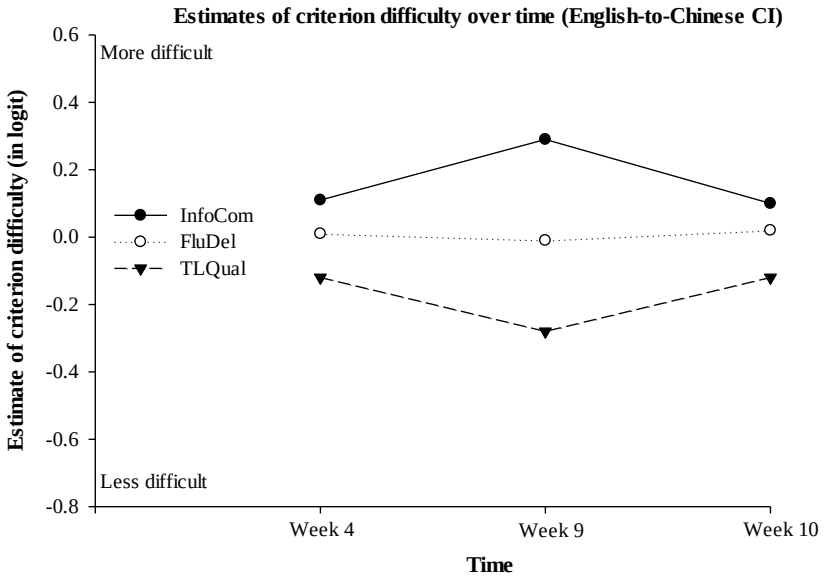

Figure 1a: Criterion or scale difficulty for English–Chinese CI

As shown in Figure 1b, for Chinese–English CI, in Week 4 the FluDel criterion or scale was the most difficult (logit = 0.26), whereas InfoCom was the easiest (logit = −0.62), which differs from the pattern observed for English–Chinese CI. The new pattern could be expressed as "FluDel > TLQual > InfoCom", which remained so across the three time points. The logit differences among the three criteria/scales were much greater than those for English–Chinese CI, ranging from 0.88 to 1.06.

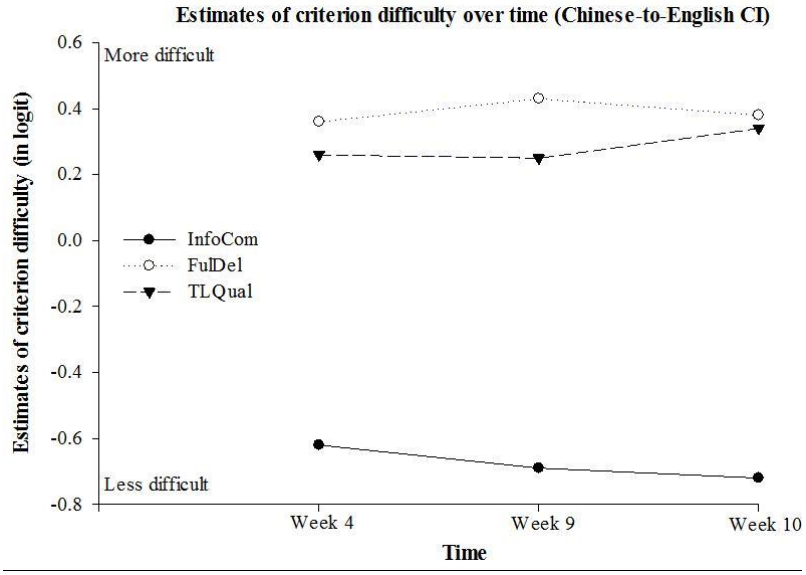

Figure 1b: Criterion or scale difficulty for Chinese–English CI

6.2. Utility of analytic rating scales

The MFRM analysis produces a series of indicators to verify the useful functioning and to diagnose the malfunctioning of rating scales. Five diagnostic indicators are highlighted in the study:

- observation count in each rating category;
- observation distribution;
- average logit measures;
- outfit statistics, and
- Rasch-Andrich threshold (Linacre, 1999).

Because of limited space, statistical indicators for the InfoCom scale are presented in Tables 2a and b. Nonetheless, the results for FluDel and TLQual were also discussed briefly.

First, Linacre (1999) recommends at least ten observations for each rating category. If the observation count for a given category is too low, Rasch calibration is imprecisely estimated and may be unstable. Inspection of Tables 2a and b reveals that observation frequency for rating category "1" was consistently below ten for Chinese–English CI. This result suggests that this category might not be particularly useful for the student profile, as it was seldom used by the raters.

Secondly, the irregular distribution of observation frequency across rating categories may indicate inconsistent category usage. Problematic are multi-modal or a roller-coaster form of distributions, and long tails of relatively infrequently used categories (Linacre, 1999). Based on the inspection of Tables 2a and b, all the frequency distributions of observation were meaningful in that they were uni-modal and peaked in central or extreme categories, except for a seemingly bi-modal distribution regarding English–Chinese CI in Week 4. However, bi-modal distributions peaking in extreme categories are also regarded as being substantially meaningful (Linacre, 1999).

Table 2a: Statistics for InfoCom rating scale in English–Chinese CI

InfoCom	Rating category	Frequency (%)	Average measure	Outfit statistics	Rasch-Andrich threshold
Week 4	1	55 (8%)	−1.43	1.1	–
	2	112 (16%)	−0.80	1.1	−1.80
	3	94 (14%)	−0.46	1.1	−0.46
	4	98 (14%)	−0.11	1.0	−0.33
	5	95 (14%)	0.22	0.8	0.06
	6	83 (12%)	0.40	1.3	0.46
	7	95 (14%)	0.80	1.0	0.49
	8	49 (7%)	1.04	1.1	1.59
Week 9	1	62 (9%)	−2.07	0.8	–
	2	154 (23%)	−1.07	1.2	−2.41
	3	150 (22%)	−0.66	1.2	−0.93
	4	114 (17%)	−0.43	1.2	−0.25
	5	80 (12%)	0.13	1.0	0.23
	6	58 (9%)	0.31	1.3	0.57
	7	47 (7%)	0.79	1.0	0.82
	8	17 (2%)	0.97	1.3	1.98
Week 10	1	25 (4%)	−1.54	1.2	–
	2	69 (10%)	−0.67	2.2	−2.26
	3	74 (11%)	−0.45	0.7	−0.79
	4	104 (15%)	−0.05	1.4	−0.64
	5	107 (16%)	0.19	1.6	0.09
	6	115 (17%)	0.72	1.5	0.48
	7	124 (18%)	1.31	1.1	0.94
	8	66 (10%)	1.63	1.3	2.18

Table 2b: Statistics for InfoCom rating scale in Chinese–English CI

InfoCom	Rating category	Frequency (%)	Average measure	Outfit statistics	Rasch-Andrich threshold
Week 4	1	7 (1%)	−1.28	0.8	–
	2	39 (6%)	−0.47	1.4	−2.56
	3	83 (12%)	−0.27	1.1	−1.23
	4	123 (18%)	−0.03	1.0	−0.53
	5	155 (23%)	0.37	0.9	−0.04
	6	130 (19%)	0.72	1.0	0.71
	7	118 (17%)	1.04	1.1	0.98
	8	28 (4%)	1.32	1.1	2.68
Week 9	1	8 (1%)	−1.70	1.3	–
	2	42 (6%)	−1.32	1.0	−3.38
	3	89 (13%)	−0.51	1.3	−1.71
	4	126 (18%)	−0.17	1.2	−0.75
	5	170 (25%)	0.33	1.2	−0.16
	6	134 (20%)	1.13	0.7	0.99
	7	82 (12%)	1.70	1.2	1.93
	8	33 (5%)	2.70	0.9	3.09
Week 10	1	7 (1%)	−1.33	1.5	–
	2	33 (5%)	−1.07	1.4	−3.12
	3	72 (11%)	−0.64	1.1	−1.73
	4	118 (17%)	−0.27	1.3	−0.97
	5	142 (21%)	0.25	1.1	−0.15
	6	138 (20%)	1.15	0.9	0.76
	7	117 (17%)	2.08	1.0	1.79
	8	57 (8%)	3.40	0.9	3.41

Thirdly, the average logit measures for rating categories should advance monotonically up the scale. In other words, observations in higher rating categories should correspond to higher logit measures. In Tables 2a and b, the average logit measures for each direction and at each time point exhibited an upward trend, consistent with the linear development of the rating categories.

Fourthly, outfit statistics (in logit) should be within the range of 0 to 2.0 (Linacre, 1999), preferably within the range 0.5 to 1.5 (Linacre, 2002). Given that the model-specified uniform value is 1.0, a rating category with a value larger than 1.5 is regarded as a *misfit* or a noise, indicating excessive variability in using the category; whereas a rating category with a value lower than 0.5 is deemed an *overfit*, suggesting too little randomness and lack of variation. Normally, a misfit is regarded as being more problematic than an overfit. For rating scales, a large outfit statistic (i.e., logit > 1.5) for a given rating category suggests unexpected usage of the category. In Tables 2a and b, for English–Chinese CI, the outfit statistic for the rating category "5" was greater than 1.5 and that for the category "2" even exceeds 2.0 in Week 10, signalling two possible misfits. Otherwise, the InfoCom scale seemed to function appropriately.

The last guideline is that Rasch-Andrich thresholds need to advance monotonically up the rating scale. Here, monotonicity indicates an increasing amount of the latent variable corresponding to a progression through the sequenced categories of the InfoCom rating scale. A review of Tables 2a and b suggests that the data complies with this guideline. Taken together, although there were some minor local problems with the InfoCom scale, on the whole the scale functioned appropriately and contributed to the measurement.

The other two scales also functioned properly overall. Regarding FluDel, only two local issues emerged for Chinese–English CI in Week 4. The observation frequency for rating category "5" (n = 5) was below 10; the Rasch-Andrich threshold value for rating category "6" did not increase as expected. As for TLQual, three minor issues surfaced: in Chinese–English CI the observation frequency for category "8" was below 10, in Week 4 (n = 9) and Week 9 (n = 7); in English–Chinese CI the frequency for category "1" was less than 10 in Week 10 (n = 8). Other diagnostic indicators for the FluDel and TLQual scales were all acceptable.

6.3. Rater behavior: Severity/leniency and self-consistency

Tables 3a and b summarize logit estimates of severity or leniency for individual raters and infit statistics of rater self-consistency. Similarly to criterion or scale difficulty, a greater logit value signals an increased degree of rater severity, accounting for the other assessment facets; a logit value of 0 indicates an impartial rater regarding their severity or leniency level. Furthermore, infit statistics are used to gauge the self-consistency of each individual rater, namely, whether a rater used rating scales consistently, holding other assessment facets constant.

In particular, like outfit statistics, the acceptable range of infit statistics (in logit) is between 0.5 and 1.5. If a rater has an infit statistic of 1.0, their rating behaviour is consistent with the Rasch model prediction (i.e., perfect fit). A rater with an infit statistic higher than 1.5 (i.e., a misfit) suggests that they display more variability in awarding scores than predicted by the model. A misfit relates to too much unexplained randomness or misinformation, which may not support useful measurement. In contrast, if a rater has an infit value lower than 0.5 (i.e., an overfit), their scores are too predictable to be informative. As always, an overfit is less problematic than a misfit.

Table 3a: Rater severity and self-consistency for English–Chinese CI

Rater behaviour	English–Chinese			
	Rater ID	Week4	Week9	Week10
Rater severity (in logit)	R01	0.23	0.34	0.45
	R02	0.35	0.31	0.34
	R03	−0.34	−0.75	−0.92
	R04	−0.31	−0.19	−0.63
	R05	0.40	0.63	0.80
	R06	−0.31	−0.33	−0.03
Self-consistency (infit statistics)	R01	0.72	0.72	0.71
	R02	0.93	1.00	1.11
	R03	*0.99*	1.33	1.04
	R04	*0.81*	0.91	0.94
	R05	*0.72*	0.77	1.21
	R06	*1.93*	1.31	1.22

Table 3b: Rater severity and self-consistency for Chinese–English CI

Rater behaviour	Chinese–English			
	Rater ID	Week4	Week9	Week10
Rater severity (in logit)	R01	0.18	0.34	0.46
	R02	0.24	0.61	0.57
	R03	−0.56	−1.51	−2.25
	R04	−0.38	−0.40	−0.70
	R05	0.59	0.67	1.03
	R06	−0.06	0.29	0.88
Self-consistency (infit statistics)	R01	0.81	0.67	0.59
	R02	0.66	0.94	1.08
	R03	0.97	1.18	1.23
	R04	0.74	0.80	0.92
	R05	0.57	0.55	0.80
	R06	2.21	1.80	1.41

For more clarity, the logit estimates of rater severity or leniency were plotted in Figures 2a and b. As can be seen in Figure 2a, for English–Chinese CI, in Week 4 three raters (i.e., Raters 01, 02 and 05) tended to be more severe, and the other three raters (i.e., Raters 03, 04 and 06) appeared to be more lenient. In Week 10, the pattern of rater severity or leniency remained.

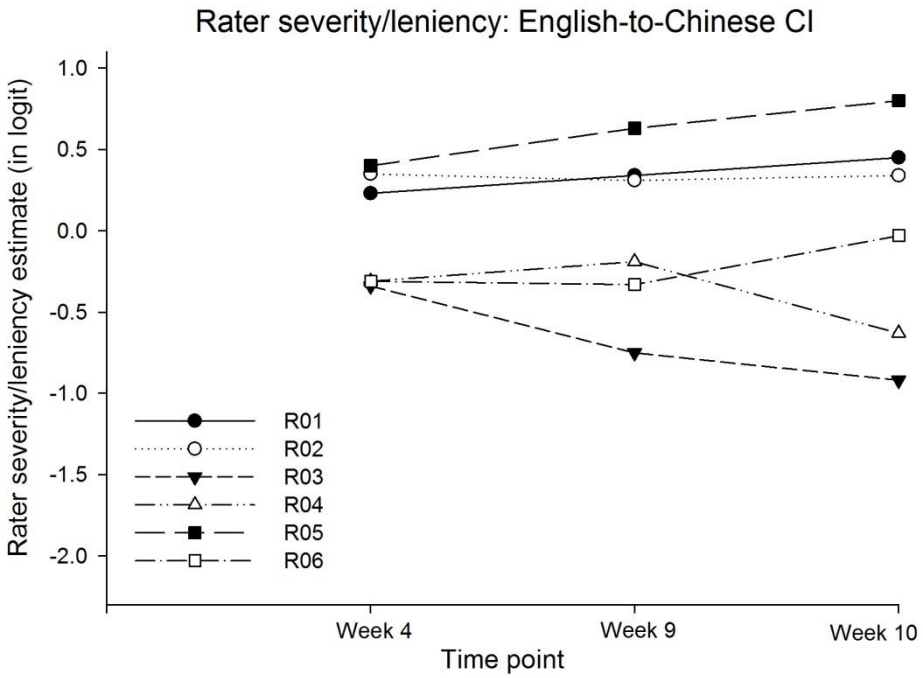

Figure 2a: Rater severity or leniency for English–Chinese CI

Specifically, three trends or findings are highlighted here. First, Raters 01 and 05 became increasingly strict over time, whereas the severity level for Rater 02 was relatively stable. In contrast, Raters 03 and 04 became more lenient by Week 10. Moreover, despite the fact that Rater 06 could be regarded as an overall lenient rater in Week 10, they actually became less lenient over time. Another finding is that in Week 10 Rater 03 became the most lenient rater, while Rater 05 became the harshest. Compared to the other raters, Raters 03 and 05 displayed the largest intra-rater changes regarding logit estimates: 0.58 and 0.4 respectively. Lastly, over time the logit estimates of rater severity or leniency remained within the range of +1.0 to −1.0; the largest logit difference was 1.72, appearing in Week 10 between Raters 03 and 05.

Figure 2b shows longitudinal changes of rater severity or leniency for Chinese–English CI. In Week 4, Raters 01, 02 and 05 appeared to be harsher, whereas Raters 03, 04 and 06 were more lenient. This pattern largely held true in Week 10, except for Rater 06, who became more severe over time (i.e., from −0.06 to 0.29 to 0.88).

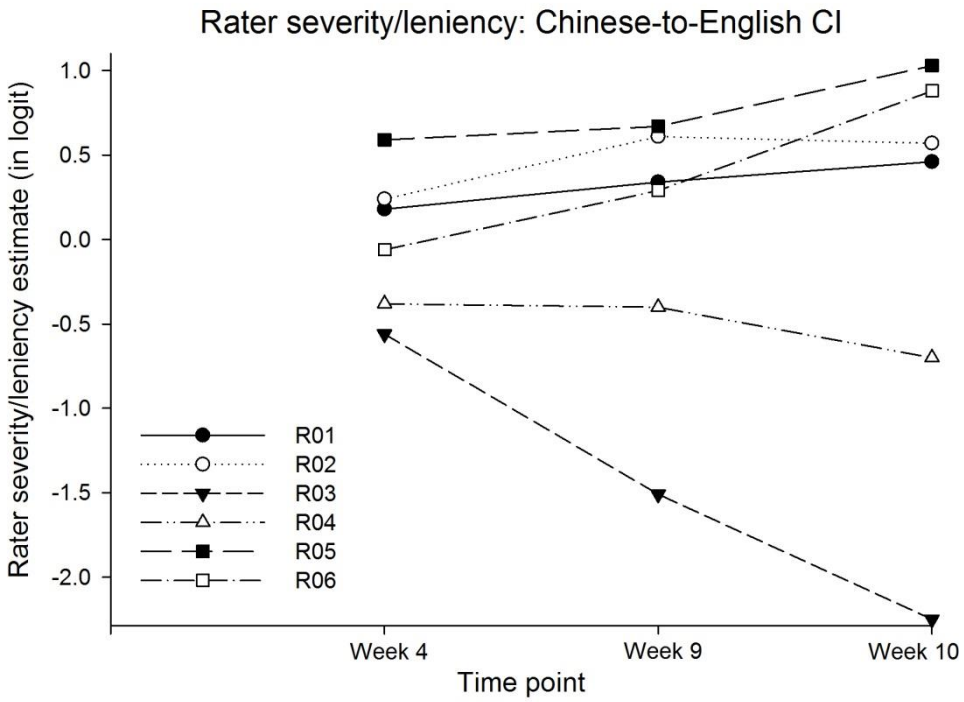

Figure 2b: Rater severity or leniency for Chinese–English CI

Specifically, and similarly to the previous findings, Raters 01, 02 and 05 started off as severe raters in Week 4 and were even more severe in Week 9 and Week 10; Raters 03 and 04 were initially lenient raters and become more lenient with the passage of time. Another similar result is that Rater 03 was the most lenient rater, while Rater 05 the harshest. This contrast is most obvious in Week 10. It also seems that the logit differences between the raters were much greater for Chinese–English CI. In particular, in Week 10 the difference was as large as 3.28 logit between Raters 03 and 05. In addition, the intra-rater difference was especially large for Rater 03 (i.e., 2.81 logit), suggesting a substantial change in rater leniency.

While a rater could be a severe or lenient rater overall, they might be self-consistent or inconsistent in their rating behaviour. To gain a better understanding of rater self-consistency or self-inconsistency over time, the infit statistics were plotted and displayed in Figures 3a and b. In Figure 3a, infit values for the majority of the raters remained within the acceptable range of 0.5 to 1.5, suggesting relatively consistent usage of rating scales for English–Chinese CI. The only exception was Rater 06. In Week 4, Rater 06 started off as a misfit (logit = 1.93), a signal of undue variability in awarding scores. However, they became more self-consistent over time and moved towards the perfect fit value of 1.0.

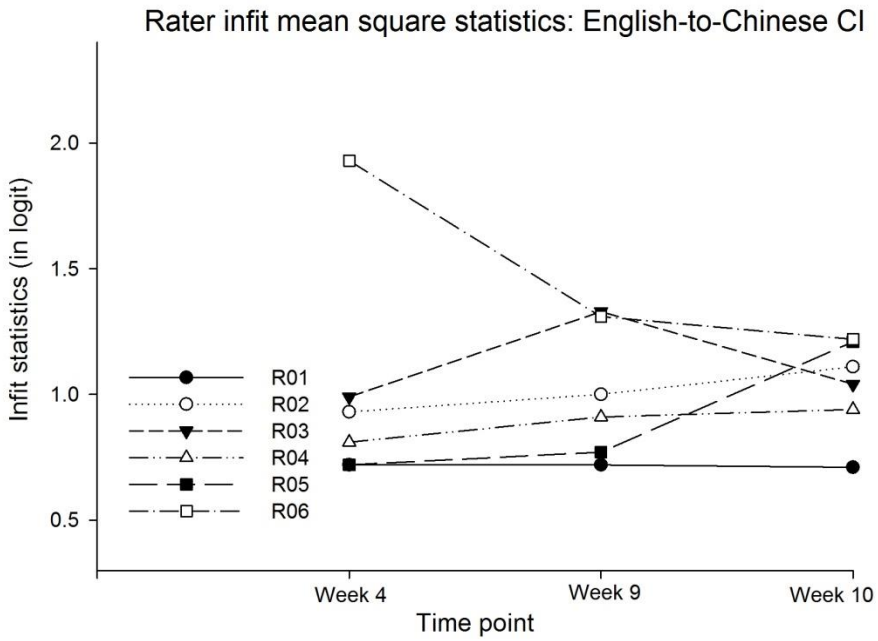

Figure 3a: Rater self-consistency for English–Chinese CI

Figure 3b shows the longitudinal developments of infit statistics for Chinese–English CI. The infit statistics for all raters (except for Rater 06) fell within the acceptable range across the three time points. With an infit value of 2.21 logit in Week 4, Rater 06 was clearly a misfit, indicating much more variability than model prediction. According to Linacre (1999), the occurrence of fit values greater than 2.0 suggests that there is more unexplained noise than explained noise, and more misinformation than information provided by the data. Rater 06 remained a misfit (1.80 logit in Week 9) until Week 10, when the infit statistic decreased to 1.41 logit.

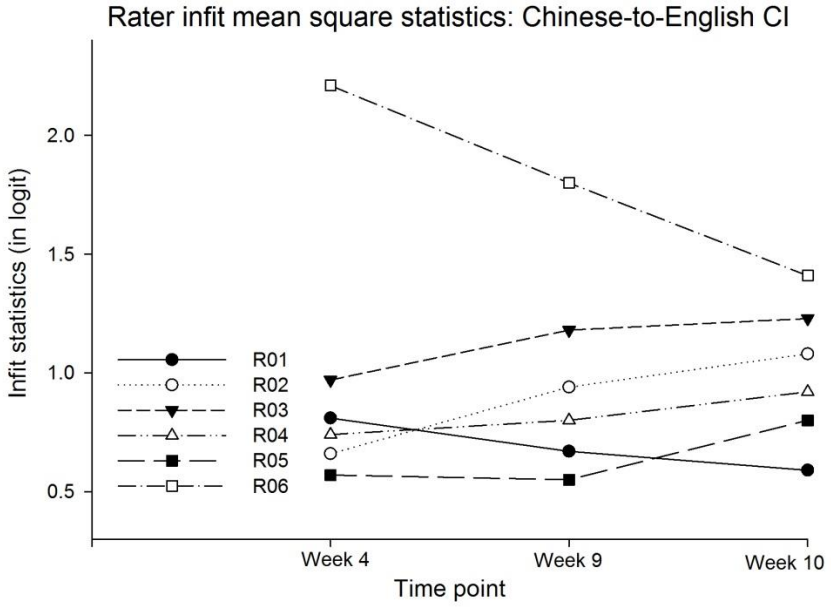

Figure 3b: Rater self-consistency for Chinese–English CI

7. Discussion

Regarding criterion or scale difficulty, the present study shows that on average for English–Chinese CI the students tended to obtain lower scores on InfoCom than FluDel and TLQual. This partially corroborates previous findings in Wu et al. (2013) and Zhao and Dong (2013). The study also extends previous research by looking into criterion or scale difficulty for Chinese–English CI: FluDel was the most difficult, followed by TLQual and InfoCom, which persisted across the three time points. These findings can be explained as follows: when interpreting English–Chinese, students who learned English as a foreign language may not comprehend the source speeches perfectly and may omit some information. However, they were able to deliver Chinese renditions smoothly with acceptable quality. When interpreting Chinese–English, although they may have a good understanding of the source speeches, they may lack TL production and reformulation resources from time to time, therefore affecting the fluency and quality of their TL renditions.

The study also contributes new empirical evidence in support of the utility of descriptor-based rating scales in assessing interpreting (see also Liu, 2013; Turner et al., 2010). Based on the five diagnostic indicators in MFRM analysis, overall the three scales functioned appropriately and effectively across time. In hindsight, the relative effectiveness may be partly attributable to the flexible nature of the scales (i.e., eight score points can be reduced to four general scale bands) and partly to the explicit instructions provided to the raters and the practical reinforcements in the training session. In particular, the raters were asked to decide first on which scale band (four choices) was most appropriate for a given instance of interpreting, and then pinpoint a score (two choices) within a band. This rating strategy decomposes into two sequential sub-processes the complex and heuristic scoring guideline that requires raters to arrive at a final score directly. The traditional scoring practice, arguably, entails multi-tasking and paralleled inferential processes in decision-making, intensifying the cognitive load on raters who have already been struggling to divide their attention between different aspects of interpreting quality (Han, 2015a; Wu, 2010).

However, the relative effectiveness of the scales should not mask a number of imperfections. Specifically, a recurring weakness is the insufficient frequency of observation (less than 10) for rating category "1" on the InfoCom scale in Chinese–English CI, and for category "8" on the TLQual scale in Chinese–English CI. From a measurement perspective, too few observation counts may jeopardize the Rasch calibration and stability of estimation (Linacre, 1999). Perhaps, more importantly, these results signal a need to focus on scale design, in particular the width of the scales (i.e., eight steps of a graded continuum) in relation to the perceived stratum of a given population. In other words, a scale may be inappropriately designed to cover a much wider spectrum of ability levels than that of the real stratum of an interested population. This could render the categories at both ends scarcely endorsed and, ultimately, make the scale less informative. In the opposite direction, the steps in a scale may be too few to encompass all the possible ability levels of a population, leading to ceiling and/or floor effects. Such design issues all boil down to the reliability and validity of scale-based measurements, which necessitates more attention and further research.

Furthermore, this study provides some useful insight into rater behaviour in the scale-based assessment of interpreting. Regarding rater severity or leniency, two interesting findings emerge from the analysis. First, the categorization of a rater as being either severe or lenient seems to be relatively stable across time for both directions. Once a rater is identified as a certain type (i.e., harsh or lenient), they tend to remain so in subsequent assessments, despite some intra-rater variation in severity or leniency estimates. In the study, it would appear that Raters 01, 02, 03, 04 and 05 conformed to this pattern. However, this explanation should not be generalized out of context; it is based only on limited observations across three time points. The other finding is that the variability among rater severity or leniency estimates for Chinese–English CI seems to be greater than that for the opposite direction. This is possibly because in assessing renditions into Chinese, the raters, who were all Chinese native speakers, may have relatively homogenous standards and might feel more comfortable in making judgements; when assessing instances of interpreting into English, the raters might have divergent quality standards and might not be able to maintain constant severity or leniency levels from one time to another. However, these explanations are preliminary, and require evidential support in future research. In addition, the statistics concerning rater self-consistency

reveal that for both directions and for each time point the raters largely remained self-consistent, with the exception of Rater 06. This is possibly because Rater 06 was a university lecturer of English–Chinese interpreting who found it difficult to adapt to the assessment methods used in the study because they had formed their own way of assessment. Although the rater training could have had some impact on improving Rater 06's rating behaviour, such impact may be limited. The design of the study (e.g., a lack of follow-up interview, a lack of pre/post-training test), however, did not allow the researcher to gain further insight into Rater 06's rating behaviour. Nevertheless, based on the self-consistency statistics, Rater 06 did become increasingly self-reliable over time, suggesting that longer exposure to and extensive use of the rating scales may help problematic raters improve rating quality over time.

Having reported the above patterns and trends, it is also appropriate to address the issue of rater variability, be it inter- or intra-rater. Currently, there is a temptation in the assessment of interpreting that requires raters to strive for absolute agreement (i.e., identical scores), either with themselves or other raters. For example, Lee (2008) and McDermid (2014) used absolute percentage agreement as an index to measure rater consensus. However, as the present study and previous research suggests (e.g., Tiselius, 2009; Wang et al., 2015), rater variability is the norm in performance assessment. Such variability may be reduced through the use of well-designed scales and rigorous rater training (Han, 2015a; Liu, 2013), but cannot be completely eliminated. At one extreme, conformist raters who give identical scores basically act like robots, repeating fully predictable outcomes. Such rating behaviour can be explained by a positivist perspective of rater-mediated assessment, which ascribes to an objective and single reality in the existential world. At the other extreme, erratic raters who generate different scores are essentially unpredictable, posing a threat to the validity of measurements. This rating behaviour could be attributed to a number of factors such as slipshod scoring methods, ill-designed rating scales and the rater's misunderstanding of assessment criteria. Both types of rater, however, do not contribute meaningful information to measurement. Perhaps, what truly matters is (inter/intra-)rater consistency, the extent to which a rater can consistently construe and use rating scales over time and the extent to which a rater can consistently rank-order performances in relation to other raters. This claim is ambitious indeed, marking a break from previous understanding of rater consistency and rater reliability in interpreting assessment. As such, empirical evidence needs to be collected to examine whether the new perspective on rater reliability is useful.

The analytic method in the study also deserves some attention. The use of MFRM to examine scale utility and rater behaviour marks a methodological departure from the traditional correlational analysis (see e.g., Lee, 2008; Liu, 2013; Turner et al., 2010). As a stochastic psychometric model, MFRM builds into its calibration variability and randomness of observed outcomes (Linacre, 2013). It is able to accommodate multiple assessment facets, create a single interval-scaled (logit) frame of reference for interpreting results, and produce diagnostic statistics for individual raters and criteria or scales. For instance, rater severity or leniency estimates were available to each rater for direct comparison, whereas in Wang et al. (2015) rater severity analysis was based on ANOVA procedures. In addition, MFRM tolerates missing data and incomplete measurement designs as long as the data are connected (Schumacker, 1999). Using Rasch measurement could enable Wang et al. (2015) to analyse all collected data. As such, MFRM analysis represents a useful and rigorous approach to analysing rater-mediated assessment of interpreting (also see Han, 2015a; Han & Slatyer, 2016).

Despite the above findings, this study has at least two limitations. Although the study provides a detailed analysis of the six raters, a group comparison as performed in Lee (2008) and Tiselius (2009) cannot be conducted due to the small rater sample size. Another criticism that could be levelled against the study is that quantitative data only were used to answer the RQs. In other similar studies (see Wang et al., 2015; Wu, 2010), qualitative data such as raters' real-time comments on rating process were collected for analysis. Wang et al. (2015), for example, identified a number of scoring techniques that may benefit further rater training. In the present study, rater-generated qualitative data, if collected, could be used to account for the patterns and irregularities emerging from the quantitative analysis.

8. Conclusion

This longitudinal study has attempted to examine raters' use of analytic rating scales to assess English–Chinese consecutive interpreting in an undergraduate-level interpreting course, based on a psychometric model called "multifaceted Rasch measurement". The analyses show that:

(a) regarding criterion or scale difficulty, InfoCom was the most difficult and TLQual the easiest in English–Chinese interpreting, while FluDel was the most difficult and InfoCom the easiest in the other direction. These patterns were also consistent over time;
(b) overall, the three rating scales functioned properly on each occasion (e.g., monotonic development of Rasch-Andrich thresholds consistent with increasing rating categories), despite some blemishes;
(c) Rater 03 was the most lenient rater, while Rater 05 the most severe; this pattern was observed in both directions and across time;
(d) the raters generally displayed greater self-consistency in assessing English–Chinese interpreting than the other way around; it would also appear that the raters, particularly Rater 06, became increasingly self-consistent over time in both directions.

Given the increasing use of analytic rating scales and the limited supporting evidence available in interpreting literature, more research is needed to demonstrate scale utility in different contexts and to ameliorate practices of scale-based assessment. Future research might profitably concentrate on: (a) evidence-based design of descriptors and rating scales that capture (para)linguistic, discoursal and pragmatic characteristics of interpreting; and (b) a longitudinal mixed-methods experimental study that examines the effectiveness of rating scales for different groups of rater receiving different interventions of rater training.

References

Angelelli, C. (2009). Using a rubric to assess translation ability: Defining the construct. In C. Angelelli & H.E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies* (pp. 13–47). Amsterdam: John Benjamins.

Barik, H. C. (1971). A description of various types of omissions, additions and errors in translation encountered in simultaneous interpretation. *Meta*, *16*(4), 199–210.

Bontempo, K., & Hutchinson, B. (2011). Striving for an 'A' grade: A case study in performance management of interpreters. *International Journal of Interpreter Education*, *3*, 56–71.

Carroll, J. B. (1966). An experiment in evaluating the quality of translations. *Mechanical Translation and Computational Linguistics*, *9*(3–4), 55–66.

Certification Commission for Healthcare Interpreters (2011). *Technical report on the development and pilot testing of the CCHI examinations*. Retrieved from http://www.cchicertification.org/images/pdfs/cchi%20technical%20report%20-%20public%20final.pdf.

Cheung, A. K. F. (2014). Anglicized numerical denominations as a coping tactic for simultaneous interpreting from English into Mandarin Chinese: An experimental study. *Forum*, *12*(1), 1–22.

Clifford, A. (2001). Discourse theory and performance-based assessment: Two tools for professional interpreting. *Meta*, *46*(2), 365–378.

Diriker, E. (2015). Conference interpreting. In F. Pöchhacker (Ed.), *Routledge encyclopedia of interpreting studies* (pp. 78–82). London: Routledge.

Gerver, D. (1969/2002). The effects of source language presentation rater on the performance of simultaneous conference interpreters. In F. Pöchhacker & M. Shlesinger (Eds.), *The interpreting studies reader* (pp. 53–66). London: Routledge.

Gile, D. (1999). Variability in the perception of fidelity in simultaneous interpretation. *Hermes*, *22*, 51–79.

Han, C. (2015a). Investigating rater severity/leniency in interpreter performance testing: A multifaceted Rasch measurement approach. *Interpreting*, *17*(2), 255–283.

Han, C. (2015b). (Para)linguistic correlates of perceived fluency in English-to-Chinese simultaneous interpretation. *International Journal of Comparative Literature & Translation Studies*, *3*(4), 32–37.

Han, C., & Slatyer, H. (2016). Test validation in interpreter certification performance testing: An argument-based approach. *Interpreting*, *18*(2), 231–258.

Han, C. (2016a). Investigating score dependability in English/Chinese interpreter certification performance testing: A generalizability theory approach. *Language Assessment Quarterly*, *13*(3), 186–201.

Han, C. (2016b). Reporting practices of rater reliability in interpreting research: A mixed-methods review of 14 journals (2004–2014). *Journal of Research Design and Statistics in Linguistics and Communication Science*, *3*(1), 49–75.

IoL Educational Trust. (2010). *Diploma in public service interpreting: Handbook for candidates*. Retrieved from http://www.iol.org.uk/qualifications/DPSI/Handbook/DPSIHB11.pdf.

Jacobson, H. E. (2009). Moving beyond words in assessing mediated interaction. In C. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies* (pp. 49–70). Amsterdam: John Benjamins.

Lee, J. (2008). Rating scales for interpreting performance assessment. *The Interpreter and Translator Trainer*, *2*(2), 165–184.

Lee, S.-B. (2015). Developing an analytic scale for assessing undergraduate students' consecutive interpreting performances. *Interpreting*, *17*(2), 226–254.

Lin, I. I., Chang, F. A., & Kuo, F. (2013). The impact of non-native accented English on rendition accuracy in simultaneous interpreting. *Translation & Interpreting*, *5*(2), 30–44.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*(2), 103–122.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2013). *A user's guide to FACETS: Program manual 3.71.2*. Retrieved from http://www.winsteps.com/a/facets-manual.pdf.

Liu, M.-H. (2013). Design and analysis of Taiwan's interpretation certification examination. In D. Tsagari & R. van Deemter (Eds.), *Assessment issues in language translation and interpreting* (pp. 163–178). Frankfurt: Peter Lang.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

McDermid, C. (2014). Cohesion in English to ASL simultaneous interpreting. *Translation & Interpreting*, *6*(1), 76–101.

Multon, K. D. (2010) Interrater reliability. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 627–629). Thousand Oaks, CA: Sage.

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

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

Schumacker, R. E. (1999). Many-facet Rasch analysis with crossed, nested and mixed designs. *Journal of Outcome Measurement*, *3*(4), 323–338.

Setton, R., & Motta, M. (2007). Syntacrobatics quality and reformulation in simultaneous-with-text. *Interpreting*, *9*(2), 199–230.

Tiselius, E. (2009). Revisiting Carroll's scales. In C. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies* (pp. 95–121). Amsterdam: John Benjamins.

Turner, B., Lai, M., & Huang, N. (2010). Error deduction and descriptors: A comparison of two methods of translation test assessment. *Translation & Interpreting*, *2*(1), 11–23.

Wang, B. H. (2011). Kou yi neng li de ping gu ji ce shi she ji zai tan – yi quan guo kou yi da sai wei li [Exploration of the assessment model and test design of interpreting competence]. *Wai yu jie*, *1*, 66–71.

Wang, J.-H., Napier, J., Goswell, D., & Carmichael, A. (2015). The design and application of rubrics to assess signed language interpreting performance. *The Interpreter and Translator Trainer*, *9*(1), 83–103.

Wu, J., Liu, M.-H., & Liao, C. (2013). Analytic scoring in interpretation test: Construct validity and the halo effect. In H.-H. Liao, T-E. Kao, & Y. Lin (Eds.), *The making of a translator: Multiple perspectives* (pp. 277–292). Taipei: Bookman.

Wu, S. C. (2010). *Assessing simultaneous interpreting: A study on test reliability and examiners' assessment behavior* (Unpublished doctoral dissertation). Newcastle University, United Kingdom.

Xi, X.-M., & Mollaun, P. (2006). *Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST*). Retrieved from https://www.ets.org/Media/Research/pdf/RR-06-07.pdf.

Zhao, N., & Dong, Y. P. (2013). Ji yu duo mian Rasch mo xing de jiao ti chuan yi ce shi xiao du yan zheng [Validation of a consecutive interpreting test based on multi-faceted Rasch model]. Jie fang jun wai guo yu xue yuan xue bao, *36*(1), 86–90.

---

1 This work was supported by the Youth Project of China's Philosophy and Social Science Fund (No. 17CYY052).

2 For a view of the descriptor-based rating scales, please visit: http://www.interpretationtestingandevaluation.com/wp-content/uploads/2016/07/Descriptor-based-analytic-rating-scales.pdf