

Raters' scoring process in assessment of interpreting: an empirical study based on eye tracking and retrospective verbalisation

Chao Han, Bingham Zheng, Mingqing Xie & Shirong Chen

To cite this article: Chao Han, Bingham Zheng, Mingqing Xie & Shirong Chen (12 Mar 2024): Raters' scoring process in assessment of interpreting: an empirical study based on eye tracking and retrospective verbalisation, The Interpreter and Translator Trainer, DOI: [10.1080/1750399X.2024.2326400](https://doi.org/10.1080/1750399X.2024.2326400)

To link to this article: <https://doi.org/10.1080/1750399X.2024.2326400>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 12 Mar 2024.



[Submit your article to this journal](#)



Article views: 709



[View related articles](#)



[View Crossmark data](#)

Raters' scoring process in assessment of interpreting: an empirical study based on eye tracking and retrospective verbalisation

Chao Han^a, Binghan Zheng^{ib}, Mingqing Xie^b and Shirong Chen^c

^aDepartment of Chinese Studies, National University of Singapore, Singapore; ^bSchool of Modern Languages and Cultures, Durham University, Durham, UK; ^cDepartment of English, Xiamen University, Xiamen, China

ABSTRACT

Human raters' assessment of interpreting is a complex process. Previous researchers have mainly relied on verbal reports to examine this process. To advance our understanding, we conducted an empirical study, collecting raters' eye-movement and retrospection data in a computerised interpreting assessment in which three groups of raters ($n = 35$) used an analytic rubric to assess 12 English-to-Chinese consecutive interpretations. We examined how the raters interacted with the source text, the rating scale, and the audio player displayed on the computer screen when they were assessing. We found that a) the source text and the rating scale were competing for the raters' visual attention, with the former attracting more attention than the latter across the rater groups; b) when the raters were consulting the rating scale, they fixated less frequently on the sub-scale of target language quality than the other two sub-scales; c) the rater groups did not seem to exhibit substantially discrepant gazing behaviours overall, although there emerged different eye-movement patterns concerning certain sub-scales; and d) the raters utilised an array of strategies and shortcuts to facilitate their assessment. We discuss these findings in relation to rater training and validation of score meaning for interpreting assessment.

ARTICLE HISTORY

Received 27 October 2022
Accepted 19 February 2024


KEYWORDS

Interpreting assessment;
rater cognition; scoring
process; eye tracking;
retrospective verbalisation

1. Introduction

Rubric-referenced, rater-mediated assessment of (spoken/signed-language) interpreting is frequently conducted in interpreter training, professional certification, and interpreting research (Han 2018).¹ Arguably, assessing interpreting quality is a complex process in which raters need to interact with multiple materials (e.g. the source text, the target text, and the scoring rubric), heed various aspects of interpreted rendition (e.g. informational, prosodic, and linguistic characteristics), and integrate their local judgements into a holistic evaluation that best captures interpreting quality (Gile 1999; Han 2022; S.-C. Wu 2010).

CONTACT Binghan Zheng  binghan.zheng@durham.ac.uk

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/1750399X.2024.2326400>

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Given the inherent complexity of interpreting assessment, empirical inquiries into raters' scoring process are of theoretical and practical significance, as an enhanced understanding can help stakeholders appreciate what numeric scores really mean (i.e. construct validity). Making sense of scores therefore requires us to contemplate such important questions as a) what aspects of interpreted rendition influence raters' decision making; b) how do raters interact with assessment materials such as the source text, the target text, and the scoring instrument (e.g. rubrics); and c) how do raters make use of a given scoring instrument. Currently, much of the previous research focuses on psychometric characteristics of rater-assigned scores (e.g. scoring reliability, accuracy, and severity) as a function of various factors including raters' professional and/or language background (Han, Hu, and Deng 2023; J. Lee 2008; Wang et al. 2015), scoring methods (Chen, Yang, and Han 2022), and assessment modalities (e.g. assessing fidelity by comparing the source text and audio-recorded interpretation versus comparing the source text and transcription of target-language rendition, see Gile 1999). Relatively little attention has been directed to the processes based on which raters evaluate interpreting and make scoring decisions. As can be seen in the review below, only a few researchers have attempted to explore the raters' scoring process in interpreting assessment (S.-B. Lee 2019; S.-C. Wu 2010, 2013), resorting largely to self-reported verbalisation data. However, the possible reactivity and non-veridicality of verbal protocols make it difficult for researchers to access and record real-time scoring processes accurately.

To broaden the boundaries of the previous research, we conducted an empirical study to investigate the scoring process of three different types of raters in a computerised interpreting assessment. To understand the scoring process, we drew on both raters' eye movements and retrospective verbalisation. Our study is expected to generate the very first set of real-time eye-movement data concerning raters' assessment of spoken-language interpreting. Overall, the study has the potential to deepen our understanding of how raters receive and evaluate interpretation and contribute to meaningful explanation of rater-assigned scores.

2. Raters' scoring process in assessment of interpreting

Overall, relatively little research has been conducted on raters' scoring process in assessment of interpreting. The prevalent approach is based on elicitation of raters' self-reported data to examine what aspects of interpreting raters heed in assessment (i.e. scoring foci) and what behaviours raters display when assessing (i.e. scoring behaviours).

One of the first empirical studies was conducted by Wu (2010, 2013), in which two groups of raters used comparative judgement to assess five English-to-Chinese simultaneous interpretations.² When making the paired comparison, the raters were asked to think aloud about their judgements. The analysis of the verbal data indicates that the raters attended to five major categories of assessment criteria: a) presentation and delivery, b) fidelity and completeness, c) audience point of view, d) interpreting skills and strategies, and e) foundational abilities for interpreting. In particular, the total number of the codes for 'presentation and delivery' and for 'fidelity and completeness' accounted for 86% of the 300 comparative judgements, indicating that these criteria attracted the most attention from the raters. In addition, concerning the scoring behaviours, the different types of raters seemed to interact with the assessment materials

differently. The interpreter raters tended to listen to the renditions and take notes, whereas the non-interpreter raters appeared to rely more on the speech script to check the fidelity of the renditions.

Another relevant study is by Wang et al. (2015), in which three raters of different professional backgrounds used a five-band, rubric-referenced analytic rating scale to evaluate sign language interpreting between English and Australian Sign Language. Data analysis reveals different behavioural patterns of scoring: the rater who was an experienced interpreter made many comments but took few notes when listening to or watching the interpretation videos; whereas the other two raters who were interpreter educators and experienced assessors commented less but made more notes. Specifically, the latter two raters commented on the interpreting performance in line with the assessment criteria, before positioning each interpretation at a specific band. They then specified a score for each of the assessment criteria and summed up the sub-scores to produce a total score for each performance. Finally, the raters verified the total score by mapping it to their overall impression and/or comparing it with the scores assigned to the previous performances.

The most relevant study, to the best of our knowledge, is by S.-B. Lee (2019), in which four raters (R1–4, i.e. interpreter trainers) assessed six undergraduate students' consecutive interpretation from English to Korean, based on their personalised holistic scoring system. The raters were asked to verbalise their thoughts when assessing. The verbal reports and their computer screen activities (i.e. keystrokes, cursor movements, and mouse clicks) were all recorded. The analysis of the multi-stream data generated an informative account for each rater's scoring process. Regarding the scoring foci, R1 and R3 focused on the content of the renditions; R2 heeded delivery; and R4 tended to consider both content and delivery and commented entirely on negative aspects of the performances. Regarding the scoring behaviours, although three out of the four raters typically assessed each performance twice, different behavioural patterns emerged. For example, in the first listening, R1 evaluated each performance sentence by sentence and awarded a provisional score, before she double-checked her decision in the second listening in which she re-examined the performance. However, in R2's re-assessment, she usually conducted spot-checking by listening to only a few segments of interest. In addition, R2 tended to highlight problematic areas in the source text (e.g. omissions, unjustifiable changes), which informed her scoring decisions. In R3's assessment, he first conducted error analysis of the performances and then repeated his analysis in the second listening for double-checking.

Taken together, the above studies have shed initial light on how raters conduct assessment of interpreting. Nonetheless, three areas of the previous research can be enhanced further. First, although it is found that different types of raters may display different scoring behaviours, such findings are inconclusive because of the small rater samples involved in each study. Second, although scoring instruments (e.g. rubrics) constitute an important part in the assessment practice, investigation into raters' interaction with such instruments is lacking. Third, the sole reliance on raters' self-reported data may lead to inaccurate and incomplete description of the scoring process.

3. The research questions

To address the issues outlined above, we recruited three groups of raters to assess 12 interpretations in a computerised assessment (explained in 4.4), using an analytic scoring rubric. Three materials typically found in this type of assessment – the source text, the scoring rubric, and the interpreting recordings playlisted in an audio player – were shown concurrently on the computer screen. We used eye tracking and retrospective interview to collect data on the scoring process, seeking to answer two major research questions (RQs):

RQ1: How did the raters interact with the source text, the rating scale, and the audio player in the assessment?

RQ2: How did the raters make use of the rubric-referenced analytic rating scale in assessment?

In answering each RQ, we focused on two aspects of the scoring process: scoring behaviours and scoring foci. To understand the raters' scoring behaviours, we examined the eye-tracking data to explore raters' eye-movement patterns during the assessment. We were particularly interested in analysing whether the three groups of raters would differ in terms of their eye movements or gazing behaviours when interacting with the three assessment materials. In parallel to the eye-movement analysis, we also examined and coded the raters' retrospections to identify specific scoring behaviours and strategies. Moreover, we drew on the rater's retrospective verbal data to examine what aspects of interpretations raters had heeded.

4. Method

4.1. Interpreting recordings

In a previous study (Chen, Yang, and Han 2022), we asked 10 experienced raters to assess a sample of 28 English-to-Chinese interpretations produced by student interpreters, based on a four-band, eight-point analytic rating scale featuring three scoring criteria of information completeness (InfoCom), fluency of delivery (FluDel), and target language quality (TLQual) (see Figures 1 and 2).³ A many-facet Rasch measurement analysis was conducted to compute fair scores (i.e. scores that are adjusted for rater severity and criterion difficulty). Based on the Rasch-calibrated scores, we purposively selected 12 interpretations comprising three distinct levels of quality (i.e. high, medium, and low), with each level having four interpretations (i.e. High – H₀₂₄, H₁₀₃, H₁₃₈, H₁₄₆; Medium – M₀₇₁, M₀₉₅, M₁₃₃, M₁₅₀; Low – L₀₁₄, L₀₅₉, L₀₆₃, L₁₃₁). The main reason for this selection is to ensure that raters are exposed to a wide range of performances so that diverse scoring behaviours could be elicited. One additional recording of medium quality (i.e. M₁₀₈) was selected for the trial scoring session. The average length of the 12 selected recordings was 2.63 minutes with a standard deviation of 0.43.

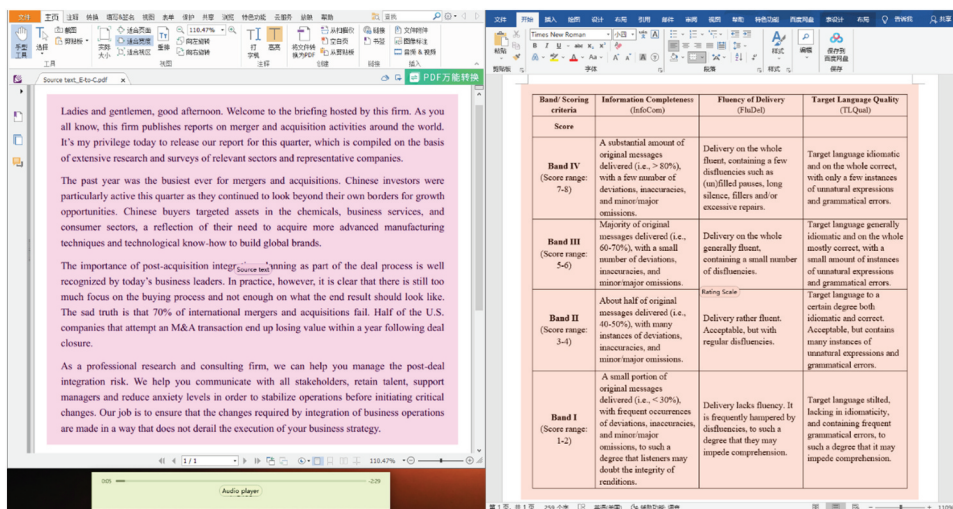


Figure 1. The presentation of the three materials/AOLs on the computer screen.

4.2. Raters

We recruited three groups of raters characterised by different capabilities and experiences. Group A consisted of 14 raters (male: $n = 2$, female: $n = 12$; age: $M = 33.9$ years, $SD = 5.2$, $Max. = 44$, $Min. = 28$) who had obtained postgraduate-level interpreting degrees, taught interpreting full-time in universities for an average of 7.5 years, assessed students' interpreting performance on a regular basis (ranging from weekly to semesterly), and were also certified interpreters (i.e. China Accreditation Testing for Translators and Interpreters Level II). Group B comprised 17 postgraduate interpreting students from Master of Translation and Interpreting programmes in major Chinese universities (male: $n = 3$, female: $n = 14$; age: $M = 24.2$ years, $SD = 1.6$, $Max. = 30$, $Min. = 23$). Group C included six postgraduate interpreting students from a major Chinese university (male: $n = 1$, female: $n = 5$; age: $M = 23.2$ years, $SD = 0.4$, $Max. = 24$, $Min. = 23$). However, what made the last group of student raters different from the previous two groups is that the six raters in Group C had used the analytic rating scale previously (see Figure 2) to assess some 60 interpretations of different source speeches. On average, the raters in Group A were more capable as interpreters and generally more experienced as assessors than those in Groups B and C, whereas the raters in Group C had obtained more experience of using the rating scale than those in Groups A and B. All raters had Mandarin Chinese as their L1 and English as their L2.

4.3. Analytic rating scale

We asked the raters to assess the interpretations using the same four-band, eight-point analytic rating scale as in Chen, Yang, and Han (2022) (see Figure 2). We used this analytic scale in the study for three reasons: a) using the same scale as in Chen, Yang, and Han (2022) would enable us to cross-validate our current assessment results by the three

Band/ Scoring criteria	Information Completeness (InfoCom)	Fluency of Delivery (FluDel)	Target Language Quality (TLQual)
Band IV (Score range: 7-8)	A substantial amount of original messages delivered (i.e., > 80%), with a few number of deviations, inaccuracies, and minor/major omissions.	Delivery on the whole fluent, containing a few disfluencies such as (un)filled pauses, long silence, fillers and/or excessive repairs.	Target language idiomatic and on the whole correct, with only a few instances of unnatural expressions and grammatical errors.
Band III (Score range: 5-6)	Majority of original messages delivered (i.e., 60-70%), with a small number of deviations, inaccuracies, and minor/major omissions.	Delivery on the whole generally fluent, containing a small number of disfluencies.	Target language generally idiomatic and on the whole mostly correct, with a small amount of instances of unnatural expressions and grammatical errors.
Band II (Score range: 3-4)	About half of original messages delivered (i.e., 40-50%), with many instances of deviations, inaccuracies, and minor/major omissions.	Delivery rather fluent. Acceptable, but with regular disfluencies.	Target language to a certain degree both idiomatic and correct. Acceptable, but contains many instances of unnatural expressions and grammatical errors.
Band I (Score range: 1-2)	A small portion of original messages delivered (i.e., < 30%), with frequent occurrences of deviations, inaccuracies, and minor/major omissions, to such a degree that listeners may doubt the integrity of renditions.	Delivery lacks fluency. It is frequently hampered by disfluencies, to such a degree that they may impede comprehension.	Target language stilted, lacking in idiomaticity, and containing frequent grammatical errors, to such a degree that it may impede comprehension.

Figure 2. The four scale-related AOIs.

rater groups; b) previous research has examined how raters used holistic scoring in interpreting assessment (S.-B. Lee 2019; S.-C. Wu 2010), but less is known about analytic rubric scoring; and c) we were interested in exploring how raters would use and interact with the analytic scale and its sub-scales.

4.4. Computerised interpreting assessment

We operationalised the interpreting assessment by giving the raters an opportunity to consult three materials – the source text, the rating scale, and the interpreting recordings – all being presented on the computer screen (i.e. a 23.8-inch EIZO FlexScan EV2451 monitor). As shown in Figure 1, the computer screen was segmented into three areas of interest (AOIs), together accounting for 72.72% of the total screen area: a) the PDF source text indicated by the pink rectangle, b) the rating scale in a Word

document indicated by the orange rectangle, and c) the interpreting recordings playlisted in a randomised order in the iTunes player, indicated by the yellow rectangle. The raters needed to write down final scores, using the keyboard, for each rubric criterion in the Word document.

This set-up resembles what usually takes place when a teacher or a student evaluates interpreting on their computer (see also S.-B. Lee 2019). In reality, alternative configurations are also possible, for instance, raters' having a hard copy of the source text, listening to interpretations, and recording final scores in an electronic file on the computer. However, this type of configuration does not allow for accurate and complete recording of raters' eye movements on a particular material and from one material to another, as at least one piece of the trio (i.e. the source text, the scoring instrument, and the recordings) is missing on the computer screen. Despite its practicality, this set-up is less favoured and indeed acknowledged as a limitation in previous research (for a detailed discussion, see Winke et al. 2015, 52). Our design – presenting the trio on the computer screen together – represents a frequent practice in real life and a potentially better design for research.

4.5. Experimental design

We had two independent variables in the experiment: a) the three rater groups and b) the AOIs we defined. To answer RQ1, we focused on the three major AOIs – the source text (480.308 cm²), the rating scale (619.581 cm²), and the audio player (36.177 cm²) (see Figure 1). To answer RQ2, we zeroed in on the four AOIs within the rating scale – Band and score (B&S, 71.319 cm²), InfoCom (133.236 cm²), FluDel (133.236 cm²), and TLQual (133.236 cm²), which represented the four columns in the rating scale from left to right (indicated by different colours in Figure 2).

Our dependent variables concerning eye movements were three common fixation-based measures: a) total number of fixations, b) total duration of fixations, and c) average duration of fixations, with the first two indices being used as indicators of raters' visual attention whereas the latter as a measure of cognitive effort (Table 1).

4.6. Procedure

The experiment consisted of several phases, as shown in Figure 3. One day before the experiment, the raters received a package of relevant materials, including the source text,

Table 1. Eye-tracking metrics.

Metric	Operational definition	Corresponding activities	Uses in previous research
Total number of fixations	The total count of all fixations in an AOI	Allocation of visual attention	Attention allocation in rater-mediated assessment (e.g. Ma and Li 2020; Ma and Winke 2022)
Total duration of fixations	The total duration of all fixations in an AOI	Allocation of visual attention	Attention allocation in rater-mediated assessment (e.g. Ma and Li 2020; Ma and Winke 2022)
Average fixation durations	The average duration of all fixations in an AOI	Cognitive effort of processing	Cognitive effort of translation-based consultation behaviour (e.g. Cui and Zheng, 2021) and sight translation (e.g. Su and Li, 2021)

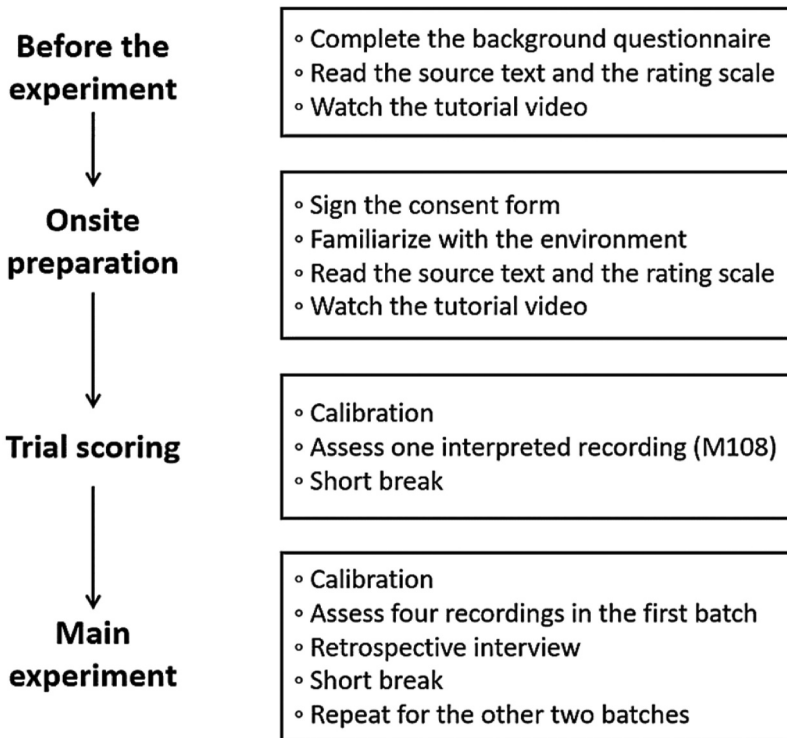


Figure 3. The experimental procedure.

the rating scale, and a six-minute tutorial video that briefly explains the scale. They were asked to familiarise themselves with the materials and watch the video at least once before the experiment, and to complete a background questionnaire (e.g. age, sex, educational background).

In the lab on the day of the study, the raters were first introduced to the experimental procedures and were informed that they could withdraw from the study at any time without any consequences. They were then seated about 65 cm from the computer monitor and re-familiarised with the source text, the rating scale, and the tutorial video. At this stage, they could mark and highlight the source text in PDF and use the highlighted PDF in the subsequent scoring session.⁴ Their eye-movements were calibrated, using a nine-point calibration method, on the eye tracker – Tobii Pro Spectrum 600 Hz. We asked the raters to self-pace their assessment. That is, they could (re)play, pause, and spot-check the recordings at will, but they had to finish their assessment before starting the next one. Additionally, before each scoring, the raters were instructed to position the rating scale appropriately and were advised not to make adjustments to it while scoring. The raters trial-assessed one recording (i.e. M_{108}) for warm-up before the operational scoring.

In the main experiment, the raters assessed a total of 12 recordings assigned to three consecutive batches, with each batch containing four recordings. Immediately after their assessment of each batch, we conducted a retrospective interview (in Chinese) with each rater to elicit answers to three questions: 1) how did you use the source text to assist

assessment; 2) how did you use the rating scale to assist assessment; and 3) how did you evaluate the last recording in each batch.

To ensure that the last recordings in the three batches were consistent across the raters, we randomly selected one sample from each set of the high-, medium-, and low-quality recordings (i.e. H₁₄₆, M₀₇₁, H₀₅₉), and then randomly placed one of the three selected recordings in the last position for each batch. That is, each rater assessed the three interpretations as the last recordings in the three batches, but in a randomised order. In addition, for each batch the first three recordings were randomly selected (with no duplicates) from the remaining nine recordings (i.e. H₀₂₄, H₁₀₃, H₁₃₈, M₀₉₅, M₁₃₃, M₁₅₀, L₀₁₄, L₀₆₃, L₁₃₁) and then randomly sequenced for each rater. Ultimately, each rater assessed the 12 recordings in a unique sequence to avoid the potential order effect.

Between two consecutive batches, the raters were given a short break and their eye-movements were calibrated each time before assessment. The whole experiment took about 90 minutes to complete and was conducted individually for each rater. The study was approved by the research ethics committee of Durham University.

4.7. Data analysis

Before the main data analysis, we examined the reliability and validity of the rater-assigned scores, by computing intraclass correlation coefficients (ICC – rater consistency) for each scoring criterion for each rater group and by correlating the rater-assigned scores from each group with the previous assessment results in Chen, Yang, and Han (2022).⁵ The ICCs were used as evidence of reliability, whereas the Pearson's correlation coefficients were interpreted as evidence of concurrent validity. We found that the ICCs (single measure) were all above .80 (except for Group B on the TLQual, ICC = .74),⁶ meaning that the raters were relatively reliable in their assessment. Pearson's correlations between the current and the previous sets of scores were above .95, pointing to very high levels of concurrent validity.

We used the Tobii Pro Lab 1.162 to process the eye-movement data. For example, fixations were identified through the Tobii I-VT filter. Apart from the AOIs, we created the times of interest (TOIs) by marking out each scoring session in which a given rater assessed an interpreting recording. A TOI normally began and ended as the raters started playing a (new) recording.⁷ In total, there were 12 TOIs for each rater, corresponding to their assessment of 12 recordings. A preliminary quality check enabled us to exclude the eye-movement data from two student raters owing to insufficient data quality (i.e. gaze sample to fixation percentage lower than 75%, see Hvelplund 2014), therefore giving us 35 valid raters.

To answer the RQs, we analysed the eye-movement data via (generalised) linear mixed models (G/LMM), using the *lme4* package (version 1.1–27.1, Bates, Kliegl, et al. 2015) in *R* (R Core Team 2021) with *p* values being computed by the *lmerTest* package (version 3.1–3, Kuznetsova, Brockhoff, and Christensen 2017). For a detailed description of how we analysed and modelled the eye-tracking data,⁸ see Appendix A.

To complement the quantitative statistical results, we analysed the raters' retrospective data from the first and the second questions.⁹ The first author (Coder A) read the transcripts of the raters' responses to identify an initial set of topics and themes, which

were also verified by the fourth author (Coder B). Based on the coding scheme, Coders A and B analysed the qualitative data independently. The inter-coder agreement, measured by percent agreement index, was 93.6% for the first interview question and 98.0% for the second interview question meaning that nine out of 10 coding decisions were the same across the two coders. The coding discrepancies, mostly pertaining to the scoring behaviours and strategies identified, were resolved through discussion.

5. Results

5.1. The raters' gazing behaviour in the three major AOIs

Descriptive statistics (M and SD) of the three eye-tracking measures for the three major AOIs are presented in [Appendix B](#). Based on the generalised mixed-effects model analyses, we find the following results. As can be seen in [Figure 4](#), regarding the total number of fixations, there was a significant main effect of the AOI, $\chi^2 = 378.628$, $df = 2$, $p < .001$. The raters fixated more frequently on the source text than on the rating scale, $z = 8.149$, $p < .001$, and more frequently on the rating scale than the audio player, $z = 14.861$, $p < .001$. There was no significant main effect of the rater group, $\chi^2 = 5.083$,

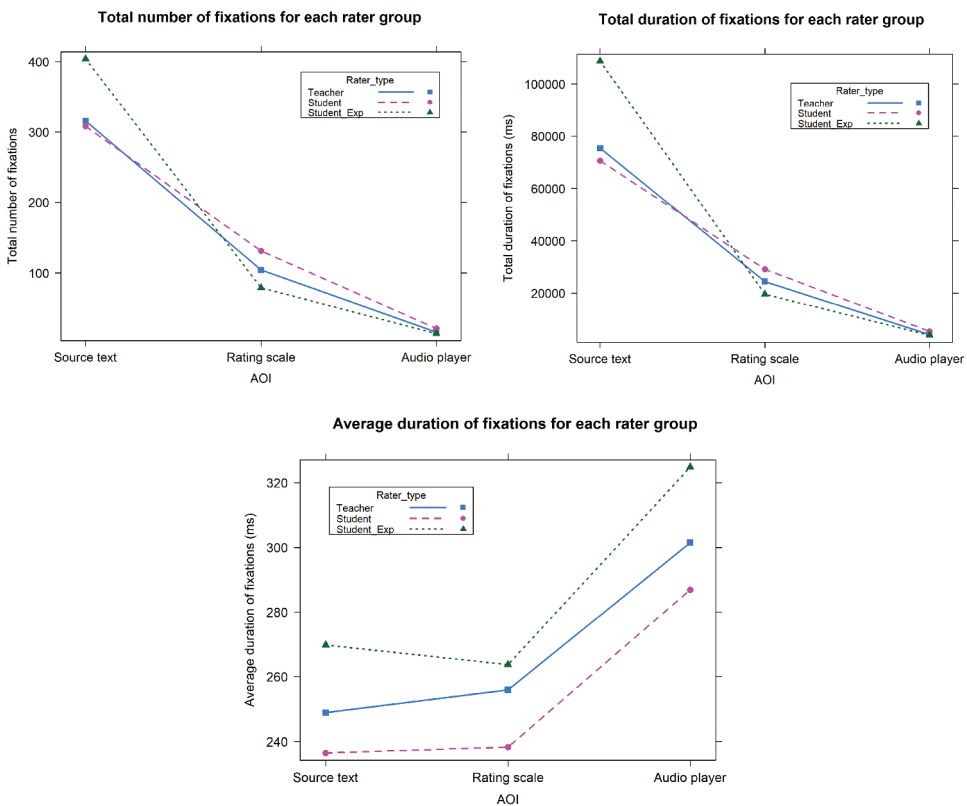


Figure 4. The pattern of the three eye-tracking measures for the three major AOIs (based on the model results).

$df = 2$, $p = .079$, and no significant interaction between the AOI and the rater group, $\chi^2 = 8.677$, $df = 4$, $p = .070$.

Regarding the total duration of fixations, the results were similar to what has been observed for the total number of fixations: a significant main effect was observed for the AOI, $F(2, 32.442) = 127.370$, $p < .001$, with the raters fixating significantly longer on the source text than on the rating scale, $t = 7.989$, $p < .001$, and longer on the rating scale than the audio player, $t = 12.361$, $p < .001$. No significant main effect of the rater group, $F(2, 32.162) = .905$, $p = .415$, and no significant interaction between the AOI and the rater group, $F(2, 32.078) = 1.890$, $p = .136$, were observed.

Regarding the average duration of fixations, there was a significant main effect of the AOI, $F(2, 32.464) = 18.642$, $p < .001$. The raters' average fixation duration for the rating scale was similar to that for the source text, $t = .217$, $p = .829$, but was significantly shorter than that for the audio player, $t = -5.992$, $p < .001$. There was no significant main effect of the rater group, $F(2, 32.069) = 2.033$, $p = .148$, and no significant interaction between the AOI and the rater group, $F(4, 32.305) = .290$, $p = .883$.

Finally, we correlated the total number of fixations and the total duration of fixations between the three major AOIs, as shown in Table 2. One noticeable pattern is the negative, moderately strong correlations of total number of fixations (Pearson's $r = -.48$, $p = .003$) and total duration of fixations (Pearson's $r = -.49$, $p = .003$) between the source text and the rating scale.

5.2. The raters' gazing behaviour in the four scale-related AOIs

Descriptive statistics (M and SD) of the three eye-tracking measures for the four scale-related AOIs are presented in Appendix C. Based on the (generalised) mixed-effects model analyses, we find the following results. As can be seen in Figure 5, regarding the total number of fixations, there was a significant main effect of the AOI, $\chi^2 = 74.406$, $df = 3$, $p < .001$. On average, the raters fixated more frequently on the subscale of InfoCom than on that of FluDel, TLQual, and B&S. There was also a main effect of the rater group, $\chi^2 = 6.120$, $df = 2$, $p = .047$: the student raters without experience had more fixations than the student raters with experience ($z = 2.413$, $p = .016$), especially on B&S and InfoCom subscales. The interaction effect between AOI and rater group was found to be non-significant, $\chi^2 = 11.170$, $df = 6$, $p = .069$.

Regarding the total duration of fixations, there was a significant main effect of the AOI, $F(3, 1618.00) = 70.863$, $p < .001$. Overall, the raters fixated longer on the sub-scale of InfoCom than on that of FluDel, TLQual, and B&S. Although there was no significant

Table 2. Inter-correlations of the fixation measures between the three major AOIs.

	Source text	Rating scale	Audio player
Total number of fixations			
Source text	1	-.48**	-.21
Rating scale		1	.41
Audio player			1
Total duration of fixations			
Source text	1	-.49**	-.23
Rating scale		1	.33
Audio player			1

** $p < .008$ (Bonferroni corrected p value = .05/6).

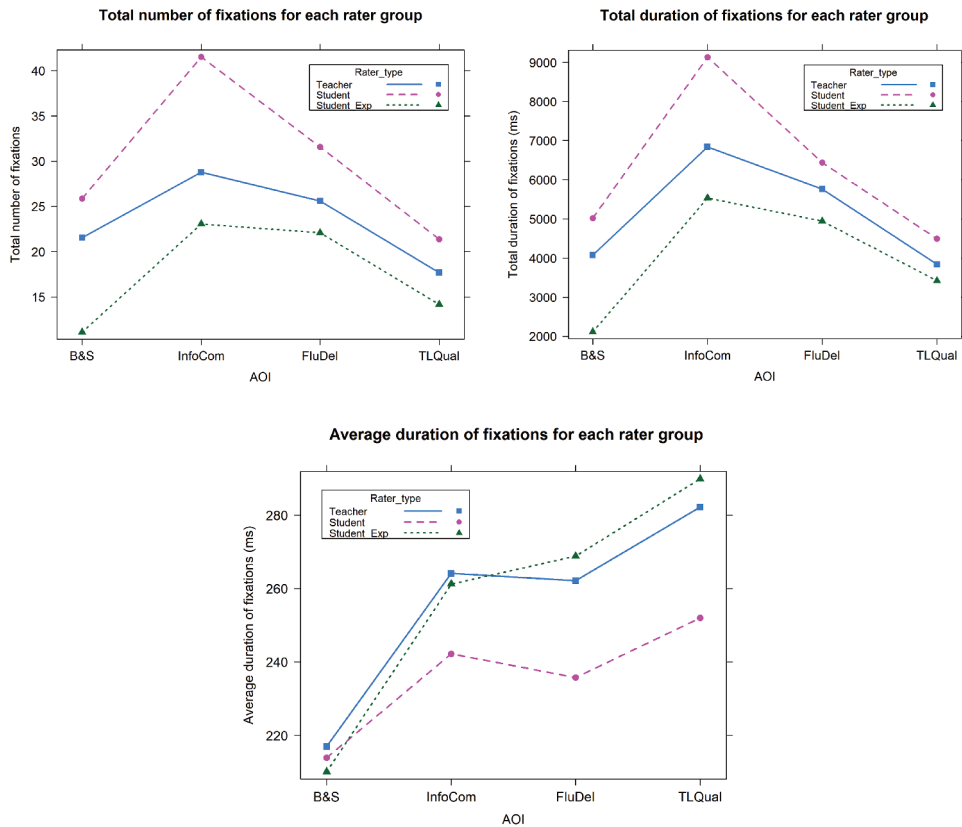


Figure 5. The pattern of the three eye-tracking measures for the four scale-related AOIs (based on the model results).

main effect of the rater group, $F(2, 32.00) = 1.973, p = .156$, the interaction effect between the AOI and the rater group was statistically significant, $F(6, 1618.00) = 4.431, p < .001$. *Post hoc* analysis revealed that the student raters without scoring experience ($t = 3.351, p = .005$) and the teacher raters ($t = 2.524, p = .047$) had more fixations on B&S than the student raters with experience.

Regarding the average duration of fixations, there was a significant main effect of the AOI, $F(3, 1571.62) = 105.996, p < .001$. Results showed that the raters' average fixation duration on B&S was shortest and that their fixations on the TLQual subscale was longest. There was no significant main effect of the rater group, $F(2, 31.99) = 1.779, p = .185$, although a significant interaction effect between the AOI and the rater group was observed, $F(6, 1571.44) = 5.663, p < .001$. *Post hoc* analysis revealed that the student raters without scoring experience had shorter average fixation duration than the teacher raters when viewing the TLQual subscale, $t = -2.552, p = .045$.

Finally, we correlated the total number and duration of fixations between the four scale-related AOIs to explore the inter-AOI fixation patterns, as can be seen in Table 3. Analysis of the correlation matrices shows that a) the positive inter-AOI correlations had moderate-to-strong relationships, except for the correlations

Table 3. Inter-correlations of the total number of fixations and the total duration of fixations between the four scale-related AOIs.

	Band & Score	InfoCom	FluDel	TLQual
Total number of fixations				
Band & Score	1	.68**	.51**	.31
InfoCom		1	.84**	.55**
FluDel			1	.80**
TLQual				1
Total duration of fixations				
Band & Score	1	.60**	.45	.25
InfoCom		1	.80**	.53**
FluDel			1	.79**
TLQual				1

** $p < .004$ (Bonferroni corrected p value = .05/12).

between B&S and the TLQual sub-scale; b) the inter-AOI correlations were consistently stronger for the adjacent AOIs than for the non-adjacent AOIs; c) the inter-AOI correlations declined in proportion to the increasing distance between the AOIs; and d) the adjacent correlations between the three sub-scales were strongest, hovering around $r = .80$.

5.3. The raters' retrospection data

5.3.1. Codes concerning the source text

Our analysis of the raters' responses to the first interview question reveals four major themes, which relate to raters' motivation behind reading the source text, specific aspects of the source text they examined, scoring behaviours and strategies, and other commentaries (see [Appendix D](#)). Although various factors may have prompted the raters to examine the source text, they specifically mentioned three reasons: a) when they were suspecting potential problems in the target renditions (e.g. relating mostly to informational discrepancy between the target information delivered and the source information stored in the raters' memory); b) when they were trying to focus on fidelity and accuracy of the target renditions; and c) when they were unable to recall the content in the source text. In addition, when the raters consulted the source text, they were most likely to focus on such specific areas as key information/words, logic, numbers, listed items, items prone to errors, and syntactically complicated sentences (see [Appendix D](#)).

Furthermore, we identified several scoring behaviours and strategies the raters utilised. One frequent behaviour is that the raters tended to listen to the interpreted recordings while reading the source text in a mostly linear fashion with the aim to examining the fidelity of the renditions (i.e. Linearity – Reading while listening, Quote 1). Another frequently mentioned behaviour relates to the sequential evaluation from the global to the local aspects of interpretation (i.e. Global-to-local evaluation, Quote 2). There are a few other interesting scoring strategies such as: a) interim scoring – assigning tentative bands/scores after examining the initial segments of interpretation (Quote 3); b) comparative evaluation – judging the quality of a rendition by comparing it to previous renditions (Quote 4); c) item-based evaluation – judging the quality by focusing on highlighted areas or items in a rendition (Quote 5); and d) strategic ignoring – ceasing to read the source text or listen to the renditions when a given performance

was considered very poor. In addition, a few comments were related to scoring granularity which pertains to the level or the unit of analysis in real-time scoring (Quote 6). Moreover, in the category of ‘Others’, we identified such strategies as listening to the recordings twice, conducting error analysis/deduction, and weighting the scoring criteria.

Quote 1: I listened to the recordings while following the source text. Most of the time, I read the source text in a linear fashion. (Rater 37)

Quote 2: I analysed whether the overall message was rendered appropriately. Based on this analysis, I then checked whether the order and details of the information were correctly interpreted. (Rater 18)

Quote 3: Generally, I evaluated the fidelity and language quality after listening to the first segment. If there were numerous fillers and incorrect syntactic segmentation, or there were frequent occurrences of certain phenomena, it was sufficient evidence to assign a band. (Rater 25)

Quote 4: The performance differences between the students would affect (my) scoring strategy. If a student’s performance was not typical, I tended to rely on the previous student’s performance as a reference for my evaluation. (Rater 21)

Quote 5: As I highlighted the key points in the source text (during the preparation stage), I focused on assessing any inconsistencies between my highlighted content and the interpreters’ renditions. (Rater 04)

Quote 6: Regarding one student’s performance, the unit of analysis was finer-grained. I assigned a score to each of the four interpreted segments, and then averaged the four scores afterwards. (Rater 21)

Finally, a large proportion of the raters (i.e. nearly 70%, $n = 23$) commented on their familiarity with the source text. Some raters expressed their familiarity because of the pre-experiment self-preparation and training, whereas others observed that the scoring experience gained during the experiment had helped them become more familiar with the source text.

5.3.2. Codes concerning the rating scale

Four main themes emerged from our analysis of the responses to the second interview question, including raters’ motivation behind using the rating scale, components of the rating scale they focused on, scoring behaviours and strategies, and other commentaries (see [Appendix E.](#)). There are four main reasons behind the raters’ utilisation of the rating scale: a) when they were to make evaluation or assign scores; b) when they were trying to gain a better understanding of the scalar descriptors; c) when they were feeling uncertain or lacking confidence in their judgements; and d) when they were spotting disfluencies in a rendition.

In addition, regarding what aspects of the rating scale the raters focused on, the retrospection data indicate that they made more comments on InfoCom, followed by FluDel, TLQual, and B&S. Specifically, the raters tended to focus on such aspects of the scale as proportional estimates (e.g. ‘> 80%’), qualifiers (e.g. ‘a small number of’, ‘regular’, ‘a few’), and key words/concepts (e.g. ‘idiomaticity’, ‘unnatural expression’).

Regarding the scoring behaviours and strategies, the raters once again mentioned several strategies similar to those identified previously, including comparative evaluation, interim scoring, and error analysis/deduction. Quite a few behaviours and strategies specifically related to the rating scale, for example, heeding a specific criterion, sequential scoring, deliberating on bands, considering final scores, revising scores, and not using the scale. Here, we highlight three strategies: a) listening to target rendition from beginning to end, b) sequential scoring, and c) not using the scale. The first strategy was mentioned by the raters when they were assessing InfoCom; the second strategy relates to a sequential pattern of decision making, referring to the process in which the raters first decided on the band, then consulted the descriptors, and finally assigned a specific score; and the third strategy was used when the raters were assessing fluency and language quality or very poor performances.

Furthermore, we identified two types of interesting comments relating to the scale. The first type mostly concerns the raters' remarks on their increasing familiarity with the rating scale and the concomitant less attention to it (Quote 7). The second type is associated with the difficulty or ease of making scoring decisions, especially when assessing students with different performance profiles (Quote 8), when providing relatively low scores (Quote 9), and when evaluating the best or the worst performances (Quote 10).

Quote 7: With more experience, I no longer needed to consult the rating scale. Nonetheless, I still read the source text (Rater 02)

Quote 8: I was struggling to assign the lowest or the second lowest band to the last three students. The language quality of their renditions was fine, but there were lots of repetitions which influenced my judgement of fluency. (Rater 32)

Quote 9: It troubled me psychologically, when I had to assign the score of three or four. This was because these scores were relatively low, and because the students had 'interpreting anxiety'. Based on my (teaching) experience, students were more susceptible to negative teacher feedback. (Rater 37)

Quote 10: It was easiest to assign scores to the best and the worst performances. But it was really difficult to score those mid-range interpretations which were a mixture of good and bad renditions. (Rater 07)

6. Discussion

Regarding RQ1 that concerns the three major AOIs, our analysis of the fixations shows a significantly higher number and longer duration of fixations on the source text than the rating scale across the rater groups. This pattern of attentional distribution may indicate that on average the raters found the source text more important, informative, and relevant to the assessment than the rating scale. Furthermore, there appears to be a trade-off between the source text and the rating scale in terms of their capability of attracting the raters' visual attention, as indicated by the moderately strong negative correlation of fixation count and fixation duration between these two AOIs ($r = -.48, -.49, p < .008$). We offer two tentative explanations to account for these patterns. One explanation is that in each assessment the raters relied mostly on the source text to examine the fidelity of

the interpreted renditions, which entails the raters' sequential reading of the source text paced by the interpreting recordings (see Quote 1), and therefore allocated less attention to the scoring *per se* which is largely based on using the rating scale. This means that the assessment is mostly characterised by the raters' comparison of the source and the target text rather than by consulting the rating scale to assign scores, although the two processes are inter-related. The other explanation is that as the rating scale became increasingly internalised by the raters, they chose not to look at the scale any more in some instances and instead focused entirely on reading the source text (see Quote 7). By contrast, given that each of the renditions was different (in terms of lexical and syntactic choices, and of the original information correctly rendered), the raters needed to consult the source text regularly for fidelity analysis, despite their growing familiarity with the source text.

Apart from the analysis of the two raw fixation measures, we further examined the average fixation duration to infer the raters' cognitive effort. One surprising result is that, overall, the raters' average fixation duration was significantly longer for the audio player than for the source text and the rating scale, suggesting relatively higher cognitive effort involved. Our tentative explanation is that the raters' interaction with the audio player involved precise visual and manual coordination as they may choose to (re)play, pause, monitor the progress of the recordings, and click and drag the slider to a specific spot in the timeline, whereas their interaction with the source text and the rating scale pertained largely to reading and analysis, requiring similar levels of processing effort.

Moreover, the mixed-effects model analyses show no statistically significant effect of the rater groups regarding the three fixation indices, meaning that, across the three major AOIs, the three rater groups displayed similar reading behaviours. In other words, on the whole rater background did not influence raters' overall visual attention distribution and cognitive effort concerning the three major AOIs. This result is not surprising, considering that all raters assessed the same 12 recordings (though in a different order), based on the same screen set-up (Figure 1). In addition, the three major AOIs were broadly defined, precluding us from obtaining more nuanced insights into their gazing behaviours.

Regarding RQ2 that pertains to the four scale-related AOIs, the mixed-effects model analysis shows that, in terms of fixation count and duration, the raters fixated significantly more and longer on the sub-scale of InfoCom than that of FluDel, TLQual, and B&S. This pattern of visual attention distribution appears to be corroborated by the frequency of raters' comments on the four scale-related AOIs (see Appendix E.), with the InfoCom sub-scale attracting more comments than the others. We offer two possibilities to (partly) explain this fixation pattern. The first possibility is that the raters fixated more and longer on InfoCom, because fidelity has long been considered the dominant quality criterion in interpreting assessment (see Gile 1999; Han 2018, 2022; J. Lee 2008). The second possibility is that when the raters were consulting the three sub-scales, they were influenced by the acquired habit of reading from left to right (from InfoCom to TLQual). The amount of visual attention may have decreased from the first (leftmost) to the last (rightmost) sub-scale. The reason why the raters heeded B&S the least is probably because there were much fewer texts associated with it than with the three sub-scales.

Based on the statistical analysis, we also find that the student raters without experience fixated more frequently and much longer than the student raters with scoring experience on the AOI of B&S. This may be because the inexperienced student raters had more

difficulty of memorising and internalising the band-score structure (e.g. Band four subsumes scores of 7 and 8, whereas Band two comprises scores of 3 and 4), and therefore had to consult B&S more frequently to access the crucial information of the score range for a given band.

Furthermore, the statistical analysis shows that the average fixation duration was longer for TLQual than for InfoCom and FluDel, indicating that the raters' interaction with the TLQual descriptors required more cognitive effort than with those of the other sub-scales. Our postulation is that the TLQual-related descriptors may be more difficult to unpack than those of InfoCom and FluDel. For instance, the TLQual sub-scale involved such abstract concepts as idiomaticity and grammatical correctness which are difficult to pin down during the assessment, whereas the InfoCom and FluDel sub-scales included more concrete and definable phenomena such as omissions, pauses, and repairs.

Finally, the mixed-effects analysis shows the statistically significant interaction effect in which the average fixation duration was longer for the teacher raters than for the student raters without experience, suggesting that a higher level of cognitive effort was involved for the teachers. As there were no statistically significant differences between these two rater groups regarding the total number and duration of fixations (i.e. a similar level of visual attention), we suppose that the differing level of cognitive effort was because the teacher raters tended to be more concerned with TLQual and therefore processed relevant descriptors in greater depth. In contrast, the student raters without experience tended to understand the TLQual-related descriptors in a more superficial manner.

Apart from the eye-movement patterns described above, we uncovered an emerging set of scoring behaviours based on the analysis of the retrospection data. Our observation on the scoring sequence – ‘determining band levels → reading criterial descriptors → assigning specific scores’ – is largely consistent with what is reported by Wang et al. (2015). In addition, the global-to-local scoring strategy resembles R2's behavioural profile reported in S.-B. Lee (2019), in which the rater focused on the overall delivery of the target renditions in the first listening and spot-checked specific segments of interest in the second listening. Other similar scoring behaviours include error analysis/deduction, interim scoring, listening to the recordings twice, and deciding on scoring granularity (see also S.-B. Lee 2019's description of the raters' behavioural profile, pp. - 259–263). One scoring behaviour unreported previously is the synchronous reading and listening, whereby the raters read the source text in a mostly linear-progressive manner, predominantly paced by the external stimuli of interpreting recordings. Another interesting observation is the raters' frequent reference to the role of memory in scoring. The raters may choose (not) to consult the source text and/or the rating scale, depending on whether relevant content had eluded them or whether target renditions contradicted the source-language content memorised by the raters.

Finally, our qualitative analysis of the retrospection data generates insight into what aspects of the source text and the rating scale the raters had heeded (i.e. the foci of attention). Specifically, when reading the source text, the raters focused on key words, numbers, listed items, and syntactically difficult sentences. That is, these source-text features may have been used as points of interest by the raters to differentiate performance qualities. In addition, when consulting the rating scale, the raters focused on such scalar descriptors as percentage estimates, qualifiers, and key concepts. Above all, we

consider that raters' assessment of interpreting is an inter-lingual, multi-sensory, and multi-tasking process, which imposes linguistic, cognitive, and psychological demands on raters, and which entails raters' interaction with relevant assessment materials to arrive at scoring decisions.

What could these results mean for future assessment practice and research? Briefly, we highlight three areas meriting further attention. First, our research reveals the complexity of the scoring process for interpreting assessment, which requires further investigation to understand how humans receive and evaluate interpreted rendition. Second, the research findings indicate the need of conducting rater training before operational scoring, specifically of helping raters to increase their familiarity with source texts and scoring methods. Third, for high-stakes interpreting testing, there is a need to validate the substantive meaning of rater-assigned scores, as raters may rely on different rules and routes to construe target renditions and make scoring decisions.

7. Conclusion

We conducted an empirical study to unpack raters' scoring process in a computerised interpreting assessment. Despite what we have found, we acknowledge three limitations of the study. First, only one direction of interpreting (i.e. English to Chinese) was assessed. Second, the small sample in each rater group and the unbalanced group size weakened the validity of the statistical conclusions. Third, the limited number and short duration of interpreted recordings ($n = 12$) may have elicited part of the raters' scoring strategies. Going forward, we call for four strands of research to verify, deepen, and advance our understanding of rater's scoring process in interpreting assessment: a) to replicate our results, using larger and different samples of interpreting recordings and raters; b) to mine time-stamped eye-tracking data to reveal nuanced behavioural patterns; c) to investigate raters' behavioural response to different assessment designs; and d) to model the relationship between raters' scoring process and their judgements.

Notes

1. Rubric-referenced, rater-mediated assessment refers to a type of assessment in which human raters play a critical role of evaluating performance and assigning scores, assisted by a rubric or a descriptor-based rating scale.
2. In interpreting assessment, comparative judgement refers to an evaluative process in which raters compare two renditions in a holistic manner and select the one of better quality.
3. The English source speech on business merger and acquisition lasted about 2.50 minutes and consisted of 270 words.
4. It turned out that two raters highlighted the PDF-styled source text.
5. As has been described in 4.1, ten experienced raters had previously assessed the same 12 interpretations. Their assessments were treated as a criterion measure to be correlated with the scores assigned by the three rater groups in the current study. By doing so, concurrent validity can be established.
6. Single measures of ICC refer to an index for the reliability of the ratings for one, typical, single rater, which is often contrasted with average measures of ICC, an index for the reliability of different raters averaged together.
7. The average duration for each TOI/rating was about 2.99 minutes ($SD = 0.70$).

8. Briefly, we specified fixed-effect structures based on our research questions. Concerning the specification of random-effect structures, previous literature has discussed two main modelling approaches: a) the ‘keep-it-maximal’ approach proposed by Barr et al. (2013), and b) the ‘model-selection’ approach recommended by Bates et al. (2015) and Matuschek et al. (2017). We ran statistical analysis using both approaches and the results were largely consistent. In the present study, we reported the results of the ‘model-selection’ approach, as it was associated with better goodness-of-fit statistics.
9. We did not analyse the responses from the third interview question, partly because the answers from the first two questions were most relevant to our study and therefore had the greatest potential to enrich our quantitative results, and partly because the qualitative data from the third question is too much to be fully discussed in the present article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the Project of Science and Technology Consulting Expert, Shanxi Province, China [No 105699901001].

ORCID

Binghan Zheng  <http://orcid.org/0000-0001-5302-4709>

References

- Barr, D. J., R. Levy, C. Scheepers, and H. J. Tily. 2013. “Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal.” *Journal of Memory and Language* 68 (3): 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>.
- Bates, D., R. Kliegl, S. Vasishth, and H. Baayen. 2015. “Parsimonious Mixed Models.” <https://arxiv.org/pdf/1506.04967>.
- Bates, D., M. Mächler, S. C. Walker, and S. Walker. 2015. “Fitting Linear Mixed-Effects Models Using Lme4.” *Journal of Statistical Software* 67 (1): 1–48. <https://doi.org/10.18637/jss.v067.i01>.
- Brauer, M., and J. J. Curtin. 2018. “Linear Mixed-Effects Models and the Analysis of Nonindependent Data: A Unified Framework to Analyze Categorical and Continuous Independent Variables That Vary Within-Subjects and/or Within-Items.” *Psychological Methods* 23 (3): 389–411. <https://doi.org/10.1037/met0000159>.
- Chen, J., H. Yang, and C. Han. 2022. “Holistic versus Analytic Scoring of Spoken-Language Interpreting: A Multi-Perspectival Comparative Analysis.” *Interpreter and Translator Trainer* 16 (4): 558–576. <https://doi.org/10.1080/1750399X.2022.2084667>.
- Cui, Y., and B. Zheng. 2021. “Consultation Behaviour with Online Resources in English-Chinese Translation: An Eye-Tracking, Screen-Recording and Retrospective Study.” *Perspectives Studies in Translation Theory and Practice* 29 (5): 740–760. <https://doi.org/10.1080/0907676X.2020.1760899>.
- Fox, J., and S. Weisberg. 2019. *An R Companion to Applied Regression*. 3rd ed. Thousand Oaks: Sage.
- Gile, D. 1999. “Variability in the Perception of Fidelity in Simultaneous Interpretation.” *Hermes: Journal of Linguistics* 22 (22): 51–79. <https://doi.org/10.7146/hjlc.v12i22.25493>.
- Han, C. 2018. “Using Rating Scales to Assess Interpretation.” *Interpreting* 20 (1): 63–101. <https://doi.org/10.1075/intp.00003.han>.

- Han, C. 2022. "Interpreting Testing and Assessment: A State-Of-The-Art Review." *Language Testing* 39 (1): 30–55. <https://doi.org/10.1177/02655322211036100>.
- Han, C., J. Hu, and Y. Deng. 2023. "Effects of Language Background and Directionality on Raters' Assessments of Spoken- Language Interpreting." *Spanish Journal of Applied Linguistics* 36 (2): 556–584. <https://doi.org/10.1075/resla.21009.han>.
- Hvelplund, K. T. 2014. "Eye Tracking and the Translation Process: Reflections on the Analysis and Interpretation of Eye-Tracking Data." *MonTI Monografías de Traducción e Interpretación* 201–223. <https://doi.org/10.6035/monti.2014.ne1.6>.
- Kuznetsova, L., P. B. Brockhoff, and R. H. B. Christensen. 2017. "LmerTest Package: Tests in Linear Mixed Effects Models." *Journal of Statistical Software* 82 (13): 1–26. <https://doi.org/10.18637/jss.v082.i13>.
- Lee, J. 2008. "Rating Scales for Interpreting Performance Assessment." *The Interpreter and Translator Trainer* 2 (2): 165–184. <https://doi.org/10.1080/1750399X.2008.10798772>.
- Lee, S.-B. 2019. "Holistic Assessment of Consecutive Interpretation: How Interpreter Trainers Rate Student Performances." *Interpreting* 21 (2): 245–269. <https://doi.org/10.1075/intp.00029.lee>.
- Ma, X., and D. Li. 2020. "翻译教师和普通读者在译文在线评阅中的认知过程研究: 基于眼动追踪数据的翻译质量评测[cognitive Processes of Translation Teachers and Ordinary Readers in Reading Translated Texts: An Eye-Tracking Study]." *Foreign Languages Research* 4:28–36.
- Matuschek, H., R. Kliegl, S. Vasishth, H. Baayen, and D. Bates. 2017. "Balancing Type I Error and Power in Linear Mixed Models." *Journal of Memory and Language* 94:305–315. <https://doi.org/10.1016/j.jml.2017.01.001>.
- Ma, W., and P. Winke. 2022. "An Investigation of the Impact of Jagged Profile on L2 Speaking Test Ratings: Evidence from Rating and Eye-Tracking Data." *Language Assessment Quarterly* 19 (4): 394–421. <https://doi.org/10.1080/15434303.2022.2078720>.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing R Version 4.1.1. <https://www.r-project.org/>.
- Su, W., and D. Li. 2021. "Exploring the Effect of Interpreting Training: Eye-Tracking English-Chinese Sight Interpreting." *Lingua* 256:103094. <https://doi.org/10.1016/j.lingua.2021.103094>.
- Wang, J., J. Napier, D. Goswell, and A. Carmichael. 2015. "The Design and Application of Rubrics to Assess Signed Language Interpreting Performance." *The Interpreter and Translator Trainer* 9 (1): 83–103. <https://doi.org/10.1080/1750399X.2015.1009261>.
- Winke, P., and H. Lim. 2015. "ESL Essay Raters' Cognitive Processes in Applying the Jacobs et Al.'s Rubric: An Eye-Movement Study." *Assessing Writing* 25 (1): 37–53. <https://doi.org/10.1016/j.asw.2015.05.002>.
- Wu, S.-C. 2010. "Assessing Simultaneous Interpreting: A Study on Test Reliability and Examiners' Assessment Behaviour." PhD Dissertation, Newcastle University.
- Wu, S.-C. 2013. "How Do We Assess Students in the Interpreting Examinations?" In *Assessment Issues in Language Translation and Interpreting*, edited by D. Tsagari and R. van Deemter, 15–33. Frankfurt: Peter Lang.
- Wu, S., D. Liu, and S. Huang. 2022. "The Effects of Over- and Under-Specified Linguistic Input on L2 Online Processing of Referring Expressions." *Journal of Psycholinguistic Research* 52 (1): 283–305. <https://doi.org/10.1007/s10936-022-09879-3>.

Appendices

Appendix A. Procedures for analysing and modelling the eye-tracking data

The AOIs and the rater groups were coded as fixed factors and compared using *successive differences contrasts* enabled by the “contr.sdif” function. These orthogonal contrasts used the grand mean as the intercept, and their associated coefficients indicated the difference between two levels of a given factor.

We log-transformed fixation durations to reduce the positive skewness. When fitting models, we started with a maximal random-effect structure that specifies the raters and the interpretations as random effects, including both random intercepts and random slopes (Barr et al. 2013). When the full model failed to converge, we trimmed down the random structure by removing the item-level correlations first, then the interactions between the fixed factors, and finally the random slopes (by first removing the factor of rater group). After the model converged successfully, we adopted a “model selection approach” (Bates, Mächler, et al. 2015) to identify a parsimonious model. We removed the random components that explain near-zero variance (Brauer and Curtin 2018) from the maximal model and built a simpler model, and then compared their AICs (i.e., Akaike Information Criterion) using Likelihood Ratio Tests (LRTs). This process stopped until all variance components differed reliably from zero. Model residuals were examined by computing their kurtosis and skewness statistics, and further visualised via quantile-quantile plots to determine whether they satisfied the normality assumption. If the model residuals violated the normality assumption, we followed Wu et al. (2022) to remove data points with standardised residuals over 2.5 standard deviations (i.e., 1.7-2.2% data loss in our case), and remodelled them.

For well-built models, we used the “Anova” function from the *lmerTest* package and the “Anova” function from the *car* package (version 3.0-11, Fox and Weisberg 2019) to extract main effects and interaction effects from LMMs (i.e., Type III Satterthwaite’s method) and GLMMs (i.e., Type III Wald chi-squared test), respectively. *Post hoc* analysis (i.e., Bonferroni-corrected pairwise comparisons) was conducted using the *emmeans* package (version 1.7.0, see <https://github.com/rvleneth/emmeans>). Finally, we combined the “allEffect” function (version 4.2-0, Fox and Weisberg 2019) from the *effects* package and the “plot” function to visualise the model results. Relevant information on the statistical models is summarised in the table below titled “Summary of the finalised statistical models”.

Summary of the finalised statistical models

Measures	Final models	Observed data	Marginal R ² (conditional R ²)
RQ1: TNF	AOI*RG+(1+AOI Part)+(1+AOI+ RG Item)	1246	0.860 (0.997)
RQ1: TDF	AOI*RG +(1+AOI Part)+(1+AOI Item)	1223	0.779 (0.874)
RQ1: AFD	AOI*RG +(1+AOI Part)	1225	0.156 (0.584)
RQ2: TNF	AOI*RG +(1+AOI Part)+(1+AOI+RG Item)	1662	0.192 (0.939)
RQ2: TDF	AOI*RG +(1 Part)	1662	0.124 (0.407)
RQ2: AFD	AOI*RG +(1 Part)+(1 Item)	1626	0.157 (0.425)

Notes: RQ = research question, TNF = total number of fixations, TDF = total duration of fixations, AFD = average fixation duration; AOI = area of interest, RG = rater group; Part = individual rater. The “Item” in our models were the target renditions or interpreting recordings.

Appendix B. Descriptive statistics of the eye-tracking measures for each major AOI

Rater type	AOI	Total number of fixations	Total duration of fixations (ms)	Average duration of fixations (ms)
Teacher (n = 14)	ST	337.67 (139.67)	85842.27 (38329.93)	252.52 (43.05)
	RS	118.97 (75.02)	29801.04 (17954.08)	259.07 (43.59)
	AP	20.15 (19.92)	6482.60 (6673.31)	331.87 (148.69)
Student (n = 15)	ST	328.29 (133.26)	79618.71 (36066.27)	240.55 (44.27)
	RS	145.54 (83.99)	35192.79 (21007.93)	240.84 (35.02)
	AP	23.62 (18.27)	6961.49 (5840.76)	299.00 (96.38)
Student_Exp (n = 6)	ST	420.37 (97.55)	112374.26 (25335.94)	272.68 (51.33)
	RS	88.88 (56.08)	23059.57 (14935.19)	263.54 (53.71)
	AP	16.13 (12.77)	5305.57 (3934.64)	338.94 (101.32)

Notes: Exp = with scoring experience; ST = Source text, RS = Rating scale, AP = Audio player; *M (SD)* = Standard deviation is presented in parenthesis.

Appendix C. Descriptive statistics of the eye-tracking measures for each scale-related AOI

Rater type	AOI	Total number of fixations	Total duration of fixations (ms)	Average duration of fixations (ms)
Teacher (n = 14)	B&S	27.08 (22.55)	6052.87 (5218.08)	221.33 (45.30)
	InfoCom	32.91 (22.74)	8480.29 (5713.88)	272.08 (77.73)
	FluDel	29.6 (21.17)	7564.36 (5322.35)	273.77 (133.05)
	TLQual	20.36 (22.60)	5642.80 (6316.94)	295.46 (111.80)
Student (n = 15)	B&S	27.08 (18.60)	6052.87 (4442.48)	221.33 (39.56)
	InfoCom	32.91 (26.66)	8480.29 (7013.81)	272.08 (53.79)
	FluDel	29.6 (30.70)	7564.36 (7727.28)	273.77 (48.17)
	TLQual	20.36 (26.24)	5642.80 (6778.48)	295.46 (77.26)
Student_Exp (n = 6)	B&S	13.33 (9.17)	2993.81 (2415.75)	216.48 (55.83)
	InfoCom	24.65 (15.74)	6639.71 (5171.52)	269.72 (77.81)
	FluDel	26.31 (24.85)	6774.85 (6353.01)	284.50 (140.20)
	TLQual	18.41 (19.46)	5024.97 (5168.39)	311.59 (149.94)

Notes: Exp = with scoring experience, B&S = Band & score, InfoCom = Information completeness, FluDel = Fluency of delivery, TLQual = Target language quality.

Appendix D. Number and percentage of the coded themes concerning the source text

Main topics/themes	Coding labels	No.	%
Motivation behind the raters' reading the source text	Suspecting problems in target rendition	18	10
	Focusing on fidelity and accuracy	15	8
	Being unable to recall source content	6	3
Specific aspects of the source text examined	Key information or words	28	15
	Logic	9	5
	Numbers	7	4
	Items prone to errors	6	3
	Enumeration or listed items	6	3
	Syntactically complicated sentences	4	2
	Scoring behaviours and strategies	Linearity – Reading while listening	15
	Global-to-local evaluation	13	7
	Interim scoring	7	4
	Comparative evaluation	5	3
	Item-based evaluation	5	3
	Strategic ignoring	4	2
	Scoring granularity	4	2
	Others	6	3
Commentaries	Familiarity with the source text	29	16
Total		187	100

Notes: No. = number, % = percentage.

Appendix E. Number and percentage of the coded themes concerning the rating scale

Main topics/themes	Coding labels	No.	%	
Motivation behind the raters' using the rating scale	Making evaluation	8	4	
	Trying to better understand descriptors	5	2	
	Feeling uncertain about one's judgement	4	2	
	Hearing disfluent rendition	4	2	
	Others	4	2	
Components of the rating scale examined	Information completeness (InfoCom)	29	13	
	Fluency of delivery (FluDel)	20	9	
	Target language quality (TLQual)	16	7	
	Band & score (B&S)	5	2	
Scoring behaviours and strategies	Comparative evaluation	19	9	
	Interim scoring	11	5	
	Not using the scale	11	5	
	Deliberating on bands	10	5	
	Sequential scoring	10	5	
	Listening to target rendition from beginning to end	6	3	
	Considering final scores	6	3	
	Heeding a specific criterion	5	2	
	Error analysis/deduction	3	1	
	Revising scores	3	1	
	Others	4	2	
	Commentaries	Familiarity with the rating scale	13	6
		Scoring ease/difficulty	13	6
Relationship between the three scoring criteria		6	3	
Total		215	100	

Notes: No. = number, % = percentage.