Interpreting quality assessment re-imagined: The synergy between human and machine scoring

Chao Han, Xiaolei Lu
Xiamen University, People’s Republic of China

Abstract
Assessment of interpreting quality is a ubiquitous social practice in the interpreting industry and academia. In this article, we focus on both psychometric and social dimensions of assessment practice, and analyse two major assessment paradigms, namely, human rater scoring and automatic machine scoring. Regarding human scoring, we describe five specific methods, including atomistic scoring, questionnaire-based scoring, multi-methods scoring, rubric scoring, and ranking, and critically analyse their respective strengths and weaknesses. In terms of automatic scoring, we highlight four assessment approaches that have been researched and operationalised in cognate disciplines and interpreting studies, including automatic assessment based on temporal variables, linguistic/surface features, machine translation metrics, and quality estimation methodology. Finally, we problematise the socio-technological tension between these two paradigms and envisage human–machine collaboration to produce psychometrically sound and socially responsible assessment. We hope that this article sparks more scholarly discussion of rater-mediated and automatic assessment of interpreting quality from a psychometric-social perspective.

Keywords
automatic machine scoring, human rater scoring, interpreting quality, interpreting studies, scoring method

Corresponding author:
Chao Han, Research Institute of Interpreting Studies, College of Foreign Languages and Cultures, Xiamen University, 422 South Siming Road, Siming, Xiamen 361005, Fujian Province, People’s Republic of China.
Email: chao.research@gmail.com
Interpreting quality assessment: A psychometric-social perspective

Interpreting quality assessment is frequently practised in the interpreting industry and academia. Assessment results often constitute crucial evidence that informs relevant stakeholders (e.g., interpreting clients/users, practitioners, educators, certifiers, researchers) of low/high-stakes decision-making in a wide range of social contexts (e.g., international conferences, community services facilities, classroom learning, interpreter testing) and for a diverse array of purposes (e.g., hiring and recruitment, admission, selection, certification). Concomitantly, assessments are capable of producing potentially far-reaching consequences, as their results may affect professional identity, personal livelihood, and social accessibility for different stakeholders. For example, assessment could determine a candidate’s admission into an interpreting programme and an interpreter’s access to a specific interpreting market. In addition, the welfare of consumers of interpreting services, particularly less advantaged groups (e.g., deaf and hard-of-hearing people), hinges partly on the quality of an assessment mechanism that is designed to certify specialised interpreters in certain practice domains. As such, while a technically or psychometrically sound assessment mechanism is a prerequisite for producing reliable and valid results, social consequences and implications are also a critical and inherent concern of both testing theorists and practitioners (see McNamara & Roever, 2006). This dimension of testing and assessment therefore relates to a larger issue of the role and impact of interpreting assessment in society.

Previous literature has touched on the social dimension of interpreting assessment by theorising consequences and impacts of assessment practice on test candidates, developers, educators, and end users of interpreting services (Campbell & Hale, 2003; Han & Slatyer, 2016; Jacobs et al., 2001; Vermeiren et al., 2009), using mostly Messick’s (1989) unitary validity theory that considers test (social) consequences as a critical aspect of test validity (i.e., consequential-aspect validity). In this article, we further address the potential socio-technological tension between human (low-tech) and machine (high-tech) scoring in interpreting assessment, and contemporary concern about technology’s potential for profound and disruptive societal transformation even in the emerging field of interpreting assessment. In contrast to the current narrative that usually describes a polarised scenario whereby human and machine scoring are placed at the ends of a continuum and pitted against each other, we argue for a synergised future in which coupling of human and machine efforts contribute to psychometrically sound, economically sensible, and socially responsible assessment in interpreting.

In the following sections, we first provide a critical psychometric-centred analysis of how interpreting quality has been assessed and quantified, based on human and machine scoring. The former represents the current dominant paradigm, while the latter pertains to an emerging paradigm that is conjectured to disrupt and replace the current practice. By doing so, we hope that our readers would be able to understand the substantive meaning underlying different measures of interpreting quality, and obtain insights into reliability, validity, and practicality of each scoring method. We then highlight tensions between human and machine scoring by describing, analysing, and problematising the dichotomy between these paradigms. Finally, we make a case for human–machine collaboration in interpreting quality assessment, and
envisage multiple scenarios in which harnessing human raters and automated scoring engines could lead to reliable and valid assessments, and create positive social consequences.

**Human rater scoring**

In all types of performance-based, rater-mediated assessment, human raters play an important role in analysing and evaluating performance. In this section, we provide a critical review of five specific scoring methods used by human raters to assess interpreting quality, including atomistic scoring, questionnaire-based scoring, multi-methods scoring, rubric scoring, and ranking, and highlight their respective psychometric properties.

**Atomistic scoring**

One of the most common methods can be referred to as atomistic scoring, in which raters focus on specific points of content (e.g., linguistic/paralinguistic features) in an interpretation, evaluate their acceptability, and tally the frequency of misinterpreted instances. Two methods that epitomise atomistic scoring are error analytic scoring and item-based or itemised scoring.

The error analytic method, also known as error analysis, has long been used in interpreting quality assessment (see Barik, 1971; Cokely, 1986; Gile, 1999; Goldman-Eisler, 1967; Setton & Motta, 2007). For instance, to assess signed-language interpretation, Cokely (1986) developed an elaborate taxonomy of interpreter miscues (e.g., omission, addition, substitution). Raters need to detect and classify miscues or errors into different categories based on their meticulous analysis of interpreted renditions against source-language input. Essentially, raters use a dichotomous scale (e.g., correct = 1, incorrect = 0) to evaluate whatever instances of rendition they consider problematic. The major benefit of the error analytic method is its capability of providing a nuanced description of faulty renditions in target-language output, which could be especially useful for pedagogical and research purposes (e.g., diagnostic feedback, analysis of a specific type of error). However, the method is reductionist, labour-intensive, time-consuming, and irreproducible (Han, 2018), which constrains large-scale application.

The item-based scoring method seems to be an enhanced version of error analysis, because the former predetermines and specifies a set of scoring units or items (i.e., lexical, syntactic, and discoursal features of source-language input) on which evaluative judgement is conducted, whereas the latter allows raters to decide on their own units of assessment (i.e., any instances of rendition that seem to be problematic). Item-based scoring is exemplified in a trilingual interpreter proficiency test involving English, Spanish, and American Sign Language (González et al., 2010), in which raters apply a dichotomous scale (e.g., correct vs. incorrect) to assess a set of preselected items. Although the itemised scoring may improve raters’ scoring consistency, concern over its validity remains. For instance, given the reductionist nature of the item-based method, assessment results provide little information on temporal and communicative aspects of interpreting such as fluency, comprehensibility, and communicative effect. Thus, itemised scoring can only partially capture the construct of interpreting quality as described in the literature.
Questionnaire-based scoring

Another method is based on questionnaires in which a checklist of assessment criteria is presented as individual questionnaire items that may be categorised into several dimensions of interpreting quality (e.g., fidelity, delivery, expression, communicative effect). When assessing interpreting performance, raters can not only record their comments on each criterion (Hartley et al., 2003; Riccardi, 1998; Schjoldager, 1995), but also provide quantitative ratings using Likert-type scales (Cheung, 2015; S.-B. Lee, 2015). For instance, S.-B. Lee’s (2015) assessment checklist includes a total of 21 criteria/items grouped into three assessment categories (i.e., content, form, and delivery) to assess Korean-to-English consecutive interpreting. In other words, raters need to evaluate 21 aspects of interpreting performance to calculate the final score for each interpretation. Similarly, in Hartley et al.’s (2003) assessment checklist, as many as 30 micro-criteria are included.

Although questionnaire-based scoring seems quick and easy, as it potentially only involves providing a rating to each criterion, the method may actually tax raters’ cognitive resources. This is because raters may find it difficult to distinguish a lengthy list of individually presented, yet oftentimes conceptually related, quality criteria, which could lead to halo effect. Raters could also give up altogether, providing indiscriminant (or even the same) ratings to conceptually related criteria. From a measurement perspective, highly correlated criteria/items may be redundant, contributing little unique information to the measurement process.

Multi-methods scoring

Multi-methods scoring has been predominantly practised by a number of interpreter certification testing agencies in the United States (Certification Commission for Healthcare Interpreters, 2012; Administrative Office of the United States Courts, 2019; PSI Services LLC, 2013). It essentially combines item-based scoring and scale-based scoring. Raters first judge the correctness of test candidates’ renditions on preselected scoring items/units (usually lexical and phrasal items), with scores of 1 and 0 indicating correct and incorrect renditions, respectively. These individual scores are then summated as an index of interpreting quality. Raters then proceed to use a holistic or an analytic Likert-type rating scale to evaluate overall interpreting performance.

In principle, multi-methods scoring overcomes the weakness associated with the itemised scoring (e.g., reductionist), by incorporating ratings from scale-based assessment. However, it is unclear how the two measures are currently integrated in a meaningful way to capture the overall quality of interpreting. In addition, if these two measures produce different assessment outcomes in terms of pass/fail, no transparent rules seem to be available to resolve such a contradiction. According to PSI Services LLC (2013), certification decisions are primarily determined by item-based scores, and scale-based ratings are rarely used to override certification decisions. It is therefore perplexing why two scoring methods are used in the first place, if measures from one method are not actually used to inform decision-making.
Rubric scoring

Rubric scoring refers to a process in which raters apply a rubric-referenced rating scale to assess interpreting performance. Particularly, detailed descriptors are created to capture typical features and characteristics for different levels along a performance continuum. Rating scales can be holistic, meaning that scalar descriptors covering different aspects of performance are associated with each performance level on the scale (Interagency Language Roundtable, 2015; Setton & Dawrant, 2016). Scales can also be analytic, with a number of sub-scales and associated descriptors targeting each major dimension of interpreting quality for each performance level (e.g., Han, 2015; J. Lee, 2008; Liu, 2013; Tiselius, 2009). Rubric scoring is increasingly used in different assessment contexts including interpreter training (J. Lee, 2008; Setton & Dawrant, 2016), professional certification (International School of Linguists, 2020; Liu, 2013; NAATI, 2019), and interpreting research (Han, 2018; Tiselius, 2009).

The often-cited benefits are its capability of producing reliable and valid assessment results, and its affordability and practicality (for a critical review, see Han, 2018). For instance, previous research has generated empirical evidence in support of scale utility in interpreting assessment, such as sound psychometric properties (Han, 2015, 2017), relatively high inter-rater reliability (J. Lee, 2008), high levels of criterion-related validity in respect to propositional analytic scoring (Liu, 2013), and more efficient use of time (Han, in press). Nevertheless, rubric scoring is not without limitations such as the difficulty of creating accurate descriptors, the resource-intensiveness involved in rater training, and potential rater effects (e.g., rater severity, halo effect, central tendency).

Ranking

The last method we highlight is ranking. Although ranking has been used to assess machine translation (MT) quality for some time (Koehn, 2010; Vilar et al., 2007), it is a relatively new concept for interpreting researchers. In general, raters need to rank-order a batch of interpretations according to their overall quality. One particular variant of the ranking method is called comparative judgement or paired comparison (Thurstone, 1927), in which judges are required to compare two like objects (in our case, two interpreted renditions) and make a binary decision about their relative qualities (i.e., deciding which rendition is of higher quality than the other). The binary outcomes from repeated comparisons between different pairs of renditions by a group of judges are then fitted to a statistical model, yielding standardised estimates of the quality of each rendition. These estimates are used to locate each rendition along a continuum of perceived quality, creating a scaled rank-order of all renditions from “worst” to “best” quality.

The method of comparative judgement is first explored by S. C. Wu (2010) to assess English-to-Chinese simultaneous interpreting produced by five postgraduate-level trainee interpreters of different abilities. Research results suggest that comparative judgement was able to reliably differentiate the students’ performances. Most recently, Han (in press) and Han and Xiao (in press) applied comparative judgement to assess spoken- and signed-language interpreting, based on an online assessment platform. Both studies found that overall, comparative judgement produced reliable and valid assessment results. However, Han (in press) evaluated the utility of comparative judgement versus rubric scoring, when
both were applied to assess the same batch of consecutive interpretations. Preliminary data analysis suggests that rubric scoring outperforms comparative judgement in terms of reliability, validity, practicality, and perceived ease of use, although the latter functioned fairly well in producing reliable and valid results.

A summary of rater-mediated assessment

Based on the review above, the scoring methods represent different ways of conceptualising interpreting quality. Some methods (e.g., error analytic scoring, item-based scoring) are based on the assumption that the global construct of “quality” can be deconstructed to many constituent parts, while others (e.g., holistic rubric scoring, comparative judgement) are grounded on the belief that quality should be examined in a more holistic manner. This part–whole relationship is best captured by Goulden’s (1992) remark that in atomistic scoring “the sum of the sub-scores for the parts is exactly equal to a valid score for the whole,” whereas in holistic assessment “the whole is not equal to the sum of the parts,” but “to the parts and their relationships” (p. 265). From the epistemological, axiological, and methodological points of view, there is no inherently superior or inferior scoring method. From a psychometric and practical point of view, however, there may be a more reliable and consistent way of assessment, because subjectivity of human judgement may be more pronounced with some scoring methods than with others.

Automatic machine scoring

In general, automatic machine scoring refers to a process wherein little human judgement is directly involved in analysing, evaluating, and quantifying interpreting quality. Importantly, automatic scoring relies on a predetermined set of algorithms that are calibrated and configured to automatically compute metrics and indices of interpreting quality. Although this paradigm has the potential to revolutionise assessment of interpreting quality, it is still at its embryonic stage and calls for more rigorous research.

In this section, we first provide an overview of important research on automatic assessment of text/speech production in applied and computational linguistics, natural language processing and MT, before reviewing recent developments regarding basic research on automatic assessment of spoken-language interpreting.

Automatic assessment of text/speech production

The overview of automatic assessment of text/speech production revolves around four lines of research in relevant fields: (a) assessment of speech fluency based on objectively measured temporal variables that can be automated by acoustic analysis software; (b) assessment of overall quality of text/speech production based on indices of linguistic/surface features that can be generated by corpus/computational linguistic tools; (c) assessment of fidelity/adequacy of MT based on algorithmic evaluation metrics; and (d) assessment of overall quality of MT based on quality estimation (QE) methodology.

Temporal variables. In language testing, researchers have investigated how utterance fluency measures relate to human raters’ perceived fluency of L1/L2 speaking (e.g., Bosker
This line of research is partly motivated by the possibility of automatically assessing perceived L1/L2 speaking fluency, given that temporal variables (e.g., pause duration, speech rate) can be measured automatically by running relevant scripts in Praat (see De Jong & Wempe, 2009). That is, if temporal variables could be closely correlated with raters’ subjective ratings on L1/L2 speaking fluency, a statistical model incorporating a combination of these variables could be built to predict human scoring.

In these studies, utterance fluency and temporal variables are objectively measured through acoustic analysis, while perceived fluency is measured by human raters/listeners. Bivariate correlation is conducted between the two sets of measures, and linear regression analysis is also performed to select the best predictors of raters’ fluency ratings. Overall, it is found that temporal variables relating to speed fluency (e.g., speech rate, mean length of runs) and breakdown fluency (e.g., pause frequency and duration) are better predictors of raters’ perceived fluency. For example, Ginther et al. (2010) analyse the spoken responses of 150 examinees in the Oral English Proficiency Test (OEPT) who represented three L1 backgrounds (Chinese, Hindi, and English), and relate temporal measures to the rater-assigned scores on the OEPT scale. They find that speech rate, speech time ratio, mean length of run, and number and length of silent pauses had strong and moderate correlations with the OEPT scale scores, and therefore propose using temporal measures of fluency when developing automatic speech scoring system.

Linguistic/surface features. In L1/L2 writing research, there has long been a tradition of examining the relationship between surface features (e.g., word count, word/sentence length, the number of grammatical errors, the type of grammatical constructions) and quality of writing samples. This sustained research effort is partly explained by the scholarly pursuit to achieve automatic writing assessment based on linguistic/surface features. The rationale behind feature-based automatic assessment is quite similar to that discussed in the previous section: human judgements of writing quality could be predicted by statistical modelling of an optimal set of linguistic/surface features automatically extracted and quantified by natural language processing tools.

Since Page (1966) argued for the use of computers to score essays in the 1970s, successive researchers have investigated automatic extraction of linguistic/surface features of writing samples to predict human evaluation of writing quality. Some pioneers (e.g., Kaplan et al., 1995, 1998; Page & Petersen, 1995) utilised grammar checking programmes (e.g., RightWriter, PowerEdit) to analyse writing samples, and built statistical models to predict human scores. Over the years, a growing number of freely available natural language processing, computational, and corpus linguistic tools has been developed and used extensively in writing assessment research. One example is Coh-Metrix (see McNamara et al., 2014), which is capable of automatically analysing writing samples and generating 100-plus indices of linguistic features relating to lexical sophistication, syntactic complexity, and cohesion. Another useful tool is L2 Syntactic Complexity Analyzer (L2SCA) (see Lu, 2010), capable of producing 14 indices of linguistic/surface features pertaining to production unit length, degree of coordination, degree of subordination, phrasal sophistication, and overall sentence complexity.
Empirical research has been conducted to identify an optimal set of linguistic predictors of human-scored writing quality (e.g., Crossley et al., 2016; Crossley & McNamara, 2012; Lu, 2017), and to examine the extent to which human judgement of writing could be predicted by automatic scoring engines (e.g., Raczyński & Cohen, 2018). Research on automatic writing assessment has attracted much attention, and contributed immensely to the design and development of automatic scoring systems such as Educational Testing Services’ e-rater® engine and Pearson’s Intelligent Essay Assessor.

Quality metrics. In MT, researchers need to evaluate MT models/systems on an ongoing basis, as such models/systems are being created, trained, tested, and modified for live work (or what is called the “production environment”). The primary purpose of automatic MT evaluation is thus to provide immediate, constant, inexpensive, and actionable feedback to improve MT models/systems. A number of metrics have been developed and used to automatically evaluate machine-translated product, including Bilingual Evaluation Understudy (BLEU, see Papineni et al., 2002), National Institute of Standards and Technology (NIST, see Doddington, 2002), Metric for Evaluation of Translation With Explicit Ordering (METEOR, see Banerjee & Lavie, 2005), and Translation Edit Rate (TER, see Snover et al., 2006).

The central idea behind BLEU, NIST, and METEOR is to compare a machine’s output with that of a human. In other words, the closer an MT output is to a professional human translation, the better its quality is supposed to be. In practice, an MT output is usually compared with multiple reference translations produced by different translators. In a nutshell, string-matching algorithmic metrics examine the extent of $n$-gram overlap for any order of $n$-grams between machine and human translation, although there are important differences among the metrics (for details, see Koehn, 2010). In addition, TER measures the minimum number of edits (e.g., insertions, deletions, substitutions, word shifts) that a human would have to perform to transform an MT output to match a higher quality reference translation exactly. By evaluating human editing effort, it is therefore able to monitor and measure MT system performance.

Pre-trained language models such as Bidirectional Encoder Representation From Transformers (BERT, see Devlin et al., 2018) are also increasingly used for automatic MT evaluation. These evaluate MT quality based on contextualised word vectors (i.e., representing words by a set of real numbers). The cosine similarity between BERT-generated word vectors in an MT output and in a human reference translation is computed as an indicator of MT quality. The higher the cosine similarity, the closer the MT output matches the human reference translation, and the better the MT quality is supposed to be (Devlin et al., 2018).

Validity of the above metrics is typically evaluated in terms of their linear correlation with human judgement of MT quality (Koehn, 2010). That is, human evaluation of the output from different MT models is considered the benchmark against which the automatic metrics are compared. From a psychometric perspective, criterion-related (concurrent) validity is of concern here. The higher the level of correlation (e.g., Pearson’s $r$), the more valid a certain metric is assumed to be. For example, Coughlin (2003) conducted an empirical study to evaluate the validity of BLEU and NIST against human judgement of MT output for multiple language pairs, and reported that the BLEU and NIST scores
closely paralleled the human judgements. However, according to Callison-Burch et al. (2006), the validity of BLEU is somewhat controversial, as the correlation between BLEU and human scores may vary from one context to another.

**QE.** Unlike MT metrics such as BLEU that calculate scores based on human reference translations, QE systems seek to predict scores without reference translation. Drawing on machine learning algorithms, QE systems are trained to make predictions on MT quality that approximate human judgement. The training is usually based on human-annotated datasets. For word-level QE tasks, each word in a training set is tagged as either “OK” or “BAD,” and then fed into a sequence-to-sequence model. Sentence-level QE tasks are usually formalised as a regression problem, in which each sentence in a dataset is annotated with Direct Assessment (typically performed by monolingual raters to assess MT adequacy) or Human-Targeted Translation Edit Rate (HTER) scores.2

As for the training methodology, current QE systems are mainly feature-based, neural-based, or pre-training-based. Feature-based QE models such as Quest++ rely heavily on the extraction of manually predefined linguistic features (Specia et al., 2013, 2015), which are then fed into the machine learning toolkit (i.e., scikit-learn) to train a traditional machine learning algorithm-based module (such as support vector regression or randomised decision trees). Overall, feature-based QE systems achieve fairly good results and are adequately interpretable.3

With the development of neural network algorithms, manually defined features are no longer state-of-the-art. Recently, researchers increasingly employ neural networks to automatically detect data patterns and extract features. In neural-based QE systems, a two-stage (predictor–estimator) neural QE model is often proposed for system training and enhancement (Kim et al., 2017). This has been shown to be highly practical and effective. However, neural-based QE systems are not flawless. They are often difficult to interpret and also require a large-scale bilingual corpus for predictor training.

Recent advances in pre-trained multilingual models (i.e., readily available models trained on large benchmark datasets) have given rise to pre-training-based QE systems. These systems utilise a self-supervised learning model and often use an mBERT-based or XLMr-based QE architecture (Ranasinghe et al., 2020; H.-J. Wu et al., 2020). There is no need to train pre-trained QE models from scratch; they simply require fine-tuning to update parameters. As a result, they are less complex, more generalisable, and require little additional bilingual data.

**Automatic assessment of interpreting: Recent research developments**

Regarding automatic assessment of interpreting, researchers have sought to emulate what has been achieved in applied linguistics, MT, and natural language processing. Among the four approaches to automatic assessment described above, analysis of temporal variables in spoken-language interpreting seems to be the most prevalent. Basically, the research methodology used by language testers (see Temporal variables, above) is extended and replicated with spoken-language interpreting. A number of researchers have examined the relationship between utterance fluency measures and human raters’ perceived fluency ratings for different modes of interpreting (Christodoulides & Lenglet, 2014; Han et al., 2020; Tohyama & Matsubara, 2006; Z.-W. Wu, 2021; Yang, 2015; Yu & Van Heuven, 2014).
For example, drawing on a relatively large sample of 320 recordings of English–Chinese consecutive interpreting, Han and Yang (submitted) correlated eight utterance fluency measures with perceived fluency ratings provided by five expert raters. They found that several temporal variables concerning speed and breakdown fluency (e.g., speech rate, phonation time ratio, mean length of unfilled pauses) registered moderate-to-strong correlations with the fluency ratings. A follow-up mini meta-analysis of six empirical studies \((n=291)\) suggests that mean length of unfilled pauses, phonation time ratio, mean length of runs, and speech rate are the influential predictors of perceived fluency. These preliminary results show that linear statistical models comprising some of these temporal variables have the potential to predict human scoring of interpreting fluency.

In terms of automatic assessment based on linguistic/surface features, we highlight two recent explorations (Liu, 2021; Ouyang et al., 2021) which conducted corpus-based computational linguistic analysis to characterise transcribed interpreting texts, and to correlate indices of linguistic features with human scoring. In Liu’s (2021) study, 64 English-to-Chinese consecutive interpreting samples were selected from the Parallel Corpus of Chinese EFL Learners (PACCEL), a corpus consisting of more than 2 million words for Chinese–English translation and interpreting data. The 64 samples were scored by two expert raters on a scale of 100 points, and were evenly distributed into four levels of performance (excellent: 100–80, good: 79–70, pass: 69–60, fail: <60). Corpus linguistic tools such as ParaConc, One Click, and Lancsbox 4.0 were used to extract a total of 41 variables related to various linguistic and paralinguistic features corresponding to three criteria of interpreting quality, namely, information accuracy, output fluency, and audience acceptability. In all, 80% of the interpreting samples \((n=52)\) were selected as the training set. To train the model, statistical modelling based on decision tree analysis was conducted to classify the interpreting samples into one of the four performance groups. The trained model was then used to automatically predict group affiliation for the remaining 20% of the interpreting samples \((n=12)\), which in turn were compared with human scores to calculate the accuracy of classification. Classification accuracy above 70% was achieved for all three quality criteria. These initial findings indicate that corpus-based profiling of interpreted texts could be further investigated as a potential avenue for automatic assessment.

In Ouyang et al.’s (2021) study, 67 Chinese-to-English consecutive interpretations sampled from the All China Interpreting Contest (ACIC) were transcribed, and then analysed by the Coh-Metrix computational tool. Of these, 50 were randomly selected to be the training set and the remaining 17 texts were used as the testing set. With the training set, the researchers correlated the Coh-Metrix indices with the actual scores assigned by the judges in the ACIC, and found that 17 indices had statistically significant correlations with the human scoring (with Pearson’s \(r\) ranging from .279 to .567, \(p<.05\)). A series of stepwise linear regressions was then conducted using the 17 indices as predictors of the human scores. The statistical model with four variables, namely, description of word count (DESWC), lexical diversity via vocd algorithm (LDVOCD), hypernymy of verbs (WRDHYPv), and density of first-person singular (WRDPRP1s), accounted for 60% of the variance in the human scores. This regression model was then applied to the testing set to predict human scores. Subsequent data analysis shows that the predicted scores had a moderately strong correlation with the actual human scores (Pearson’s \(r=.517, p<.05\)). This exploratory study is encouraging because it points to the possibility of predicting...
human raters’ quality assessments of spoken-language interpreting based purely on linguistic indices from Coh-Metrix analysis.

With respect to automatic assessment of interpreting based on MT metrics (e.g., BLEU), it seems that little relevant research has been reported in the interpreting literature. One exception, however, is Lu and Han’s (submitted) recent pilot study to explore to what extent MT metrics such as BLEU, NIST, METEOR, TER, and BERT correlate with human scores provided by different types of raters using different scoring methods. The rationale behind the exploration is that the MT metrics, which can be automatically computed, could be a useful indicator of fidelity in spoken-language interpreting, if relatively strong and statistically significant correlation exists between the MT metrics and human scores across different assessment scenarios, language pairs, and directions of interpreting.

Specifically, in Lu and Han’s (submitted) study, 56 English–Chinese consecutive interpretations were sourced from a larger corpus consisting of 320 recordings produced by 160 students at different stages of interpreting learning, with half of the 56 samples pertaining to the English-to-Chinese direction and the other half to the opposite direction. The interpretations in both directions were scored by four groups of bilingual raters (teachers, students, L1 Chinese speakers, and L1 English speakers), using three scoring methods (rubric analytic scoring, rubric holistic scoring and comparative judgement). A total of eight sets of scores were therefore available for each direction of interpreting. That is, the teacher and the student raters used all three scoring methods to assess the interpretations, thus providing six sets of scores (i.e., $2 \times 3 = 6$). These raters had Chinese as their L1, and English as L2. The other two groups of raters, namely, the Chinese-L1 and the English-L1 raters, only used a rubric-referenced, analytic rating scale to assess the interpretations, thereby generating another two sets of scores. In total, there were eight sets of scores for each interpreting direction. In addition, all interpretations were transcribed verbatim. Four interpreters/interpreting trainers generated four reference translations against which the student interpretations were compared. The programming language Python was used to compute five MT metrics, BLEU, NIST, METEOR, TER, and BERT, for each transcribed interpreting text in each direction. Statistical analysis was then conducted to examine the direction and the magnitude of correlation between the MT metrics and the human scores in different assessment scenarios. Overall, positive and moderate-to-strong correlations were found between most of the MT metrics and the human scores, indicating that it is possible to automate interpreting quality assessment based on the MT metrics.

Finally, a few researchers from the fields of computer science and natural language processing have explored automatic assessment of spoken-language interpreting using QE methodology (Le et al., 2018; Stewart et al., 2018). In Le et al.’s (2018) study, an interpreting corpus of 6,700 utterances was built, comprising Automatic Speech Recognition (ASR) output, verbatim transcript, text translation, speech translation, and post-edited translation. The corpus was then fed into several word confidence estimation systems that combine nine ASR features for speech transcription (i.e., acoustic features, graph features, linguistic features, lexical features, and context features) and 24 features related to MT (e.g., alignment context features, longest target/source N-gram length, target polysemy count). Subsequent data analysis showed that the MT features had a greater impact on the results, whereas the ASR features contributed complementary
information to interpreting assessment. The researchers propose several novel word-
level quality assessment models, and argue for the application of QE systems as a com-
plementary tool for assessing human-generated interpretations.

In Stewart et al.’s (2018) study, the feature-based QE model Quest++ was augmented
with four additional interpreting-specific features (ratio of pauses/ hesitations/ incomplete
words, ratio of non-specific words, ratio of “quasi-” cognates, and ratio of number of
words) to evaluate English–Japanese, English–French, and English–Italian simultaneous
interpreting. At the training stage, seven episodes of TED Talks taken from the Nara
Institute of Science and Technology (NAIST) TED SI corpus (Shimizu et al., 2013) and
transcripts/ translations (739 utterances for English–French and 731 utterances for
English–Italian) of the European Parliament speech drawn from the European Parliament
Translation and Interpreting Corpus (EPTIC) were fed into three models, that is, original
Quest++ (baseline), trimmed Quest++ (ablated through cross-validation), and aug-
mented Quest++, to obtain QE scores. Results show that the predicted scores of the
augmented Quest++ had statistically significant correlations with the METEOR scores
(with Pearson’s $r$ ranging from .41 to .69, $p < .05$), which outperformed the other two
models in all language pairs. An additional inspection of the data shows that even the
baseline QE system obtained respectable correlation coefficients (with Pearson’s $r$ rang-
ing from .32 to .63, $p < .1$), which supports the applicability of the QE methodology to
predict simultaneous interpreting quality.

In summary, an emerging body of empirical studies has been conducted recently to
explore various routes to achieve automatic assessment of interpreting. This multi-
pronged exploration is potentially inclusive, exciting, and inspiring, a positive and surg-
ing momentum that will probably continue in the foreseeable future.

A dichotomy between human and machine scoring?

Previous literature and research (including ours, for example, Chen & Han, 2021) seems
to distinguish human scoring from machine scoring, signalling a tendency to portray a
binary division or a dichotomy. This dichotomy seems to be magnified by extensive
media coverage that predicts disruptive societal transformation caused by technological
advances such as machine learning and artificial intelligence. Such a divisive portrait
could be problematic, because use of traditional scoring methods nowadays (e.g., rubric
scoring, ranking) is necessarily facilitated and automated, to varying degrees, by modern
technologies such as computer and statistical software, whereas machine scoring is sim-
ply impossible without human intervention. More importantly, as human scores are used
to train and condition machine scoring, human and machine scoring are related to each
other in a fundamental and critical way that cannot be severed.

A closer analysis enables us to put the respective strengths and weaknesses of human
and machine scoring into perspective, rather than focus solely on their differences. First,
human scoring tends to be more accessible and democratic than machine scoring. While
everyone can evaluate performance using traditional scoring methods, access to auto-
matic scoring systems (even if they exist) can be very restrictive because these systems
are likely to be proprietary. For example, two currently available automatic speech
assessment systems used for assessing L2 speaking, *SpeechRater* and *Versant*, are
developed by large testing organisations (the Educational Testing Service and Pearson respectively) and cannot be accessed freely by L2 English learners. Second, human scoring tends to be slow and time-consuming, whereas machine scoring is generally fast and efficient. The instantaneous availability of assessment results is important in certain assessment contexts (e.g., diagnostic assessment). Third, the usefulness of human-generated scores could be undermined by rater variability, whereas machine scoring delivers consistent and reliable measurements. Fourth, validity of human- and machine-generated scores could both be problematic. Human raters may rely on their own idiosyncratic criteria in assessment, introducing construct-irrelevant variance to their scores. This underscores the importance of rater training in the human scoring paradigm. Similarly, an automatic scoring system may be based on an inadequately designed algorithm that falls short of capturing the construct of interpreting quality, thereby producing questionable metrics. This underscores the importance of rigorous algorithm development in the automatic scoring paradigm. Finally, human raters are able to provide engaging, thought-provoking, and personalised assessment feedback, which is particularly valuable in formative and diagnostic assessment (see Fowler, 2007; Han & Fan, 2020; S.-B. Lee, 2017), whereas automatic scoring machines are considerably less flexible in generating meaningful feedback.

Based on the above analysis, we argue that it is more useful to envisage a continuum of scoring methods requiring different levels of human involvement. While we are pursuing an increasing level of automation in interpreting quality assessment per se, we should also look out for an optimal position on that continuum where human raters collaborate productively with automatic scoring systems. Human–machine collaboration in language (translation and interpreting) assessment could ease growing concerns over redundant human labour and supremacy of automatic scoring.

**Human–machine collaboration for better interpreting assessment**

The proposed continuum of scoring methods permits a potentially complementary and mutually reinforcing relationship between human raters and machine scoring engines. This is not to say that machine scoring will play its part in the immediate future in assessing interpreting quality. On a cautionary note, we envisage three possible experimental scenarios in which collaboration between human raters and machine scoring algorithms can be explored and its effectiveness evaluated.

One possible experimental exploration is classroom-based, low-stakes formative assessment in which automatic scoring can be used in conjunction with human scoring. The former could generate speedy, metrics-based feedback to students, while detailed diagnosis can be provided later by teachers. This combination could address the dilemma that educators have to choose between instantaneity and comprehensiveness when providing formative feedback (see Price et al., 2010). A positive consequence would be the opportunity for fine-tuning teachers’ feedback practice while engaging with students immediately after assessment.

A second possible experiment concerns medium-stakes summative assessment in which machine scoring is applied in tandem with human scoring to produce weighted
final scores. Based on prior knowledge of automatic and human scores (e.g., validity, reliability), different weighting schemes can be used to offset respective weaknesses. For instance, more weighting could be assigned to automatic scores, if the human raters involved are known to be inconsistent. This could lead to an efficient assessment practice and a sufficient level of scoring reliability.

A third experiment—that may involve uncertainties and spark controversies—would be assessment of interpreting quality in large-scale, high-stakes professional certification testing such as the China Accreditation Tests for Translators and Interpreters (CATTI), which attracts tens of thousands of test candidates each year, representing a gigantic workload for human raters. A possible experimental exploration is to use automatic scoring to screen performance for subsequent human scoring. That is, only those candidates whose performance scores sufficiently high on certain automated metrics will be evaluated by human raters. The threshold chosen will then be of critical importance, and should demonstrate with a high degree of confidence that the low scores assigned by the automatic system would have also been assigned by human raters. An alternative would be to use machine scoring as a quality-control mechanism for human scoring, with a third rater required to adjudicate when differences between human-assigned and automated scores surpass a pre-determined threshold. In this scenario, machine scoring essentially complements human scoring.

However, it should be noted that automatic scoring algorithms could potentially discriminate against certain interpreting features, particularly when the algorithms are not strictly modelled on human rater’s cognition and decision-making parameters. Given that pass/fail is the only concern in certification testing, decisions of “borderline pass” and “borderline fail” made by automated scorers need to be cross-validated by human raters.

Apart from potential experiments to increase assessment accuracy (i.e., psychometric robustness), human–machine collaboration may also represent a positive means to achieve a more socially responsible assessment mechanism. This is because solely relying on human raters or on automatic scoring engines may bring about undesirable social consequences. On one hand, relying only on human raters involves a huge amount of labour, incurring financial costs that will be ultimately borne by test takers, and leading to rater overwork and fatigue, and accordingly, to less psychometrically sound measurements. On the other hand, using only automated scorers to make certification decisions is likely to alter test takers’ learning and preparation strategies, especially when they are informed of the parameters involved in automatic scoring, and also to curtail test takers’ willingness to produce user-oriented interpretation, if they know that no humans listen to their renditions. Furthermore, if the algorithms involved systematically penalise the presence or lack of a specific feature that has little to do with interpreting quality, such scoring engines could potentially disadvantage some test takers and call into question the social fairness of the assessment.

**Conclusion**

Interpreting quality assessment is one of the most important topics in interpreting practice, education, certification, and research. In this article, we describe two major
interparadigms of assessing the quality of interpreting, namely, human rater scoring and automatic machine scoring. Specifically, we provide a critical review of five specific scoring methods used by human raters: atomistic scoring, questionnaire-based scoring, multimethods scoring, rubric scoring, and ranking. In addition, we provide an informative overview of the emerging paradigm of automatic machine scoring, by focusing on research in other disciplines and tracking recent developments in interpreting quality assessment. Finally, we challenge the dichotomy between the two assessment paradigms, and entertain the possibility of human–machine collaboration in future assessment practice.

Despite our informative description and critical review, we are aware of potential biases. For instance, the article does not represent a systematic review based on rigorous data sampling, and a review of relevant publications cannot be exhaustive. Nevertheless, we remain confident that our article contributes meaningfully to the current and future debate on interpreting quality assessment, by synthesising past assessment endeavours and shaping possible future directions of practice and research. Going forward, we envisage three major strands of research. One strand relates to enhancing current scoring methods by reducing human raters’ cognitive load in real-time assessment. Another strand concerns the continuous development and improvement of automatic scoring algorithms to achieve reliable and valid assessment outcomes. The last strand pertains to experimenting with human–machine collaboration with the aim of producing socially responsible assessments. We hope that this article motivates evidence-based, substantive inquiries into the use of different scoring methods, sparks scholarly interest in exploring multidisciplinary approaches to machine scoring, and provokes in-depth discussion on human–machine collaboration in assessment practice.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by China’s National Social Science Foundation (No. 18CYY010).

ORCID iDs
Chao Han https://orcid.org/0000-0002-6712-0555
Xiaolei Lu https://orcid.org/0000-0002-6929-4110

Notes
1. We are acutely aware of the increasing number of relevant metrics proposed and introduced in machine translation (MT), a field that is dynamic, fast-changing, and constantly evolving. In this article, however, we only focus on these five metrics as they represent the most popular ones among MT researchers.
2. When calculating Human-Targeted Translation Edit Rate (HTER), the MT output is evaluated against a post-edited version of itself, as opposed to being scored against existing reference translations produced by human translators.
3. As a result, they often serve as the baseline for WMT quality estimation (QE) tasks. WMT is an annual workshop/conference on machine translation, which can be traced back to 2006 (see http://www.statmt.org/wmt16/).
4. Similar research has also been conducted to explore automatic assessment of written translation (see Conde, 2011; Yuan, 2016).
5. Voctd allows the input of samples of any size, reporting D as the measure of lexical diversity.
6. A recent study by Chung (2020) examines the use of BLEU and METEOR to automatically evaluate human translation from German to Korean. The overall Pearson’s correlations between human and machine scoring were .849 and .862 for Bilingual Evaluation Understudy (BLEU) and Metric for Evaluation of Translation With Explicit Ordering (METEOR), respectively ($p < .01$).

References

Administrative Office of the United States Courts. (2019). *Federal court interpreter certification examination: Examinee handbook*. https://www.prometric.com/sites/default/files/2019-09/fcice_examineehandbook.pdf
Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72). https://www.aclweb.org/anthology/W05-0909/
Barik, H. C. (1971). A description of various types of omissions, additions and errors of translation encountered in simultaneous interpretation. *Meta*, 16(4), 199–210. https://doi.org/10.7202/001972ar
Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2012). What makes speech sound fluent? The contributions of pauses, speed and repair. *Language Testing*, 30(2), 159–175. https://doi.org/10.1177/0265532212455394
Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. https://www.aclweb.org/anthology/E06-1032.pdf
Campbell, S., & Hale, S. (2003). Translation and interpreting assessment in the context of educational measurement. In G. Anderman & M. Rogers (Eds.), *Translation today: Trends and perspectives* (pp. 205–224). Multilingual Matters.
Certification Commission for Healthcare Interpreters. (2012). *Technical report on the development and pilot testing of the Certified Healthcare Interpreter™ (CHİ™) examination for Arabic and Mandarin*. Certification Commission for Healthcare Interpreters. https://cchicsertification.org/uploads/CCHI_Technical_Report-CHI-Arabic_Mandarin.pdf
Chen, J., & Han, C. (Eds.) (2021). *Testing and assessment of interpreting: Recent developments in China*. Springer Nature. https://doi.org/10.1007/978-981-15-8554-8
Cheung, A. K. F. (2015). Scapegoating the interpreter for listeners’ dissatisfaction with their level of understanding: An experimental study. *Interpreting*, 17(1), 46–63. https://doi.org/10.1075/intp.17.1.03che
Christodoulides, G., & Lenglet, C. (2014). Prosodic correlates of perceived quality and fluency in simultaneous interpreting. In N. Campbell, D. Gibbon, & D. Hirst (Eds.), *Proceedings of the 7th speech prosody conference* (pp. 1002–1006). https://www.isca-speech.org/archive/SpeechProsody_2014/pdfs/194.pdf
Chung, H.-Y. (2020). Automatic evaluation of human translation: BLEU vs. METEROR. *Lebende Sprachen*, 63(1), 181–205. https://doi.org/10.1515.les-220-0009
Cokely, D. (1986). The effects of lag time on interpreter errors. *Sign Language Studies*, 53, 341–375. https://doi.org/10.1353/sls.1986.0025

Conde, T. (2011). Translation evaluation on the surface of texts: A preliminary analysis. *Journal of Specialised Translation*, 15, 69–86. https://www.jostrans.org/issue15/art_conde.pdf

Coughlin, D. (2003). Correlating automated and human assessments of machine translation quality. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.694.8573&rep=rep1&type=pdf

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32, 1–16. https://doi.org/10.1016/j.jslw.2016.01.003

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading*, 35(2), 115–135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

De Jong, N., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2), 385–390. https://doi.org/10.3758/BRM.41.2.385

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805

Doddington, G. (2002). Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. https://dl.acm.org/doi/pdf/10.5555/1289189.1289273

Fowler, Y. (2007). Formative assessment: Using peer and self-assessment in interpreter training. In C. Wadensjö, B. E. Dimitrova, & A-L. Nilsson (Eds.), *The critical link 4: Professionalisation of interpreting in the community* (pp. 253–262). John Benjamins. https://doi.org/10.1075/btl.70.28fow

Gile, D. (1999). Variability in the perception of fidelity in simultaneous interpretation. *Hermes*, 22, 51–79. https://doi.org/10.7146/hjlcb.v12i22.25493

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, 27(3), 379–399. https://doi.org/10.1177/0265532210364407

Goldman-Eisler, F. (1967). Segmentation of input in simultaneous interpretation. *Journal of Psycholinguistic Research*, 1, 127–140. https://doi.org/10.1007/BF01068102

González, R. D., Gatto, P., & Bichsel, J. (2010). Constructing a valid and reliable trilingual interpreting testing instrument. In R. L. McKee & J. E. Davis (Eds.), *Interpreting in multilingual, multicultural contexts* (pp. 55–82). Gallaudet University Press. https://muse.jhu.edu/chapter/272584/pdf

Goulden, N. R. (1992). Theory and vocabulary for communication assessments. *Communication Education*, 41(3), 258–269. https://doi.org/10.1080/03634529209378887

Han, C. (2015). Investigating rater severity/leniency in interpreter performance testing: A multifaceted Rasch measurement approach. *Interpreting*, 17(2), 255–283. https://doi.org/10.1075/intp.17.2.05han

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. https://lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/429/407

Han, C. (2018). Using rating scales to assess interpretation: Practices, problems and prospects. *Interpreting*, 20(2), 59–95. https://doi.org/10.1075/intp.00003.han

Han, C. (in press). Analytic rubric scoring versus comparative judgment: A comparison of two approaches to assessing spoken-language interpreting. *Meta.*
Han, C., Chen, S.-J., Fu, R.-B., & Fan, Q. (2020). Modeling the relationship between utterance fluency and raters’ perceived fluency of consecutive interpreting. *Interpreting*, 22(2), 211–237. https://doi.org/10.1075/intp.00040.han

Han, C., & Fan, Q. (2020). Using self-assessment as a formative assessment tool in an English–Chinese interpreting course: Student views and perceptions of its utility. *Perspectives*, 28(1), 109–125. https://doi.org/10.1080/0907676X.2019.1615516

Han, C., & Slatyer, H. (2016). Test validation in interpreter certification performance testing: An argument-based approach. *Interpreting*, 18(2), 231–258. https://doi.org/10.1075/intp.18.2.04han

Han, C., & Xiao, X.-Y. (in press). A comparative judgment approach to assessing Chinese sign language interpretation. *Language Testing*.

Han, C., & Yang, L.-Y. (submitted). Relating utterance fluency to perceived fluency of interpreting: A partial replication and a mini meta-analysis. Manuscript submitted for publication.

Hartley, A., Mason, I., Peng, G., & Perez, I. (2003). *Peer- and self-assessment in conference interpreter centre for languages*. Linguistics and Area Studies. Heriot Watt University.

International School of Linguists. (2020). *Diploma in public service interpreting learner handbook*. https://www.islinguists.com/wp-content/uploads/2017/07/ISL-DPSI-Handbook-v2.9.pdf

Interagency Language Roundtable. (2015). *IRL skill level descriptions for interpretation performance*. https://www.govtilr.org/Skills/interpretationSLDsapproved.htm

Jacobs, E. A., Lauderdale, D. S., Meltzer, D., Shorey, J. M., Levinson, W., & Thisted, R. A. (2001). Impact of interpreter services on delivery of health care to limited-English-proficient patients. *Journal of General Internal Medicine*, 16(7), 468–474. https://doi.org/10.1046/j.1525-1497.2001.016007468.x

Kaplan, R. M., Burstein, J., Lu, C., Rock, D., Kaplan, B., & Wolff, S. (1995). *Evaluating a prototype essay scoring procedure using off the shelf software*. Educational Testing Service.

Kaplan, R. M., Wolff, S., Burstein, J., Lu, C., Rock, D. A., & Kaplan, B. A. (1998). Scoring essays automatically using surface features (ETS Research Report 98-39). *Educational Testing Service*. https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.1998.tb01788.x

Kim, H., Lee, J.-H., & Na, S.-H. (2017). Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the 2nd conference on machine translation*. http://www.statmt.org/wmt17/pdf/WMT63.pdf

Koehn, P. (2010). *Statistical machine translation*. Cambridge University Press.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145–164. https://doi.org/10.1016/j.system.2004.01.001

Le, N.-T., Lecouteux, B., & Besacier, L. (2018). Automatic quality estimation for speech translation using joint ASR and MT features. *Machine Translation*, 32(4), 325–351. https://doi.org/10.1007/s10590-018-9218-6

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. (2015). Developing an analytic scale for assessing undergraduate students’ consecutive interpreting performances. *Interpreting*, 17(2), 226–254. https://doi.org/10.1075/intp.17.2.04lee

Lee, S.-B. (2017). University students’ experience of “scale-referenced” peer assessment for a consecutive interpreting examination. *Assessment and Evaluation in Higher Education*, 42(7), 1015–1029. https://doi.org/10.1080/02602938.2016.1223269

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). Peter Lang.
Liu, Y.-M. (2021). Exploring a corpus-based approach to assessing interpreting quality. In J. Chen & C. Han (Eds.), *Testing and assessment of interpreting: Recent developments in China* (pp. 159–178). Springer. https://doi.org/10.1007/978-981-15-8554-8_8

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing and implications for writing assessment. *Language Testing, 34*(4), 493–511. https://doi.org/10.1177/0265532217710675

Lu, X.-L., & Han, C. (submitted). *Automatic assessment of spoken-language interpreting: A multi-scenario exploratory study based on machine translation evaluation metrics*. Manuscript submitted for publication.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Blackwell.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

NAATI. (2019). *Certified interpreter test assessment rubrics*. https://www.naati.com.au/media/2245/ci_spoken_assessment_rubrics.pdf

Ouyang, L.-W., Lv, Q.-X., & Liang, J.-Y. (2021). Coh-Metrix model-based automatic assessment of interpreting quality. In J. Chen & C. Han (Eds.), *Testing and assessment of interpreting: Recent developments in China* (pp. 179–200). Springer. https://doi.org/10.1007/978-981-15-8554-8_9

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*, 238–243.

Page, E. B., & Petersen, N. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*, 561–565.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318). https://www.aclweb.org/anthology/P02-1040.pdf

Préfontaine, Y., Kormos, J., & Johnson, D. E. (2016). How do utterance measures predict raters’ perceptions of fluency in French as a second language? *Language Testing, 33*(1), 53–73.

Price, M., Handley, K., Millar, J., & O’Donovan, B. (2010). Feedback: All that effort, but what is the effect? *Assessment and Evaluation in Higher Education, 35*(3), 277–289. https://doi.org/10.1080/02602930903541007

PSI Services LLC. (2013). *Development and validation of oral examinations for medical interpreter certification: Mandarin, Russian, Cantonese, Korean, and Vietnamese forms*. https://nbcmi.memberclicks.net/assets/docs/tech-report-development-validation-language-forms.pdf

Raczynski, K., & Cohen, A. (2018). Appraising the scoring performance of automated essay scoring systems—Some additional considerations: Which essays? Which human raters? Which scores? *Applied Measurement in Education, 31*(3), 233–240. https://doi.org/10.1080/08957347.2018.1464449

Ranasinghe, T., Orasan, C., & Mitkov, R. (2020). TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th international conference on computational linguistics* (pp. 5070–5081). https://arxiv.org/pdf/2011.01536.pdf

Riccardi, A. (1998). Evaluation in interpretation: Macrocriteria and microcriteria. In E. Huang (Ed.), *Teaching translation and interpreting 4: Building Bridges* (pp. 115–127). John Benjamins. https://doi.org/10.1075/btl.42.14ric
Schjoldager, A. (1995). Assessment of simultaneous interpreting. In C. Dollerup & V. Appel (Eds.), Teaching translation and interpreting 3: New horizons (pp. 187–195). John Benjamins. https://doi.org/10.1075/btl.16.26schn

Setton, R., & Dawrant, A. (2016). Conference interpreting: A trainer’s guide. John Benjamins. https://doi.org/10.1075/btl.121

Setton, R., & Motta, M. (2007). Syntacrobaties: Quality and reformulation in simultaneous-with-text. Interpreting, 9, 199–230. https://doi.org/10.1075/intp.9.2.04set

Shimizu, H., Neubig, G., Sakti, S., Toda, T., & Nakamura, S. (2013). Constructing a speech translation system using simultaneous interpretation data. http://workshop2013.iwslt.org/downloads/Constructing_a_Speech_Translation_System_using_Simultaneous_Interpretation_Data.pdf

Snover, M., Dorr, B., Schwartz, R. M., icciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In Proceedings of the 7th conference of the Association for Machine Translation in the Americas (pp. 223–231). https://aclanthology.org/2006.amta-papers.25

Specia, L., Gustavo, P., & Scarton, C. (2015). Multi-level translation quality prediction with Quest++. In Proceedings of ACL-IJCNLP 2015 system demonstrations. https://www.aclweb.org/anthology/P15-4020.pdf

Specia, L., Shah, K., De Souza, J. G. C., Cohn, T., & Kessler, F. B. (2013). QuEst—A translation quality estimation framework. In Proceedings of the 51st annual meeting of the Association for Computational Linguistics: System demonstrations. http://www.qt21.eu/launchpad/sites/default/files/Quest.pdf

Stewart, C., Vogler, N., Hu, J.-J., Boyd-Graber, J., & Neubig, G. (2018). Automatic estimation of simultaneous interpreter performance. In The 56th annual meeting of the Association for Computational Linguistics (pp. 662–666). https://aclweb.org/anthology/P18-2105

Thurstone, L. L. (1927). A law of comparative judgment. Psychological Review, 4, 273–286. https://doi.org/10.1037/h0070288

Tiselius, E. (2009). Revisiting Carroll’s scales. In C. V. Angelelli & H. E. Jacobson (Eds.), Testing and assessment in translation and interpreting studies (pp. 95–121). John Benjamins. https://doi.org/10.1075/ata.xiv.07tis

Tohyama, H., & Matsubara, S. (2006). Influence of pause length on listeners’ impressions in simultaneous interpretation. In InterSpeech 2006 —ICSLP (pp. 893–896). http://slp.itc.nagoya-u.ac.jp/web/papers/2006/ICSLP06_tohyama.pdf

Vermeiren, H., Gucht, J. V., & De Bontridder, L. (2009). Standards as critical success factors in assessments: Certifying social interpreters in Flanders, Belgium. In C. V. Angelelli & H. E. Jacobson (Eds.), Testing and assessment in translation and interpreting studies (pp. 291–330). John Benjamins. https://doi.org/10.1075/ata.xiv.14ver

Vilar, D., Leusch, G., Ney, H., & Banchs, R. E. (2007). Human evaluation of machine translation through binary system comparisons. In Proceedings of the 2nd workshop on statistical machine translation (pp. 96–103). https://dl.acm.org/doi/10.5555/1626355.1626368

Wu, H.-J., Wang, Z.-X., Ma, Q.-S., Wen, X.-J., Wang, R.-C., Wang, X.-L., Zhang, Y.-L., Yao, Z.-P., & Peng, S.-Y. (2020). Tencent submission for WMT20 quality estimation shared task. In Proceedings of the 5th conference on machine translation (pp. 1062–1067). http://www.statmt.org/wmt20/pdf/2020.wmt-1.124.pdf

Wu, S. C. (2010). Assessing simultaneous interpreting: A study on test reliability and examiners’ assessment behavior. https://theses.ncl.ac.uk/1122/1/Wu%202011.pdf

Wu, Z.-W. (2021). Chasing the unicorn? The feasibility of automatic assessment of interpreting fluency. In J. Chen & C. Han (Eds.), Testing and assessment of interpreting: Recent developments in China (pp. 143–158). Springer. https://doi.org/10.1007/978-981-15-8554-8_7
Yang, L.-Y. (2015). An exploratory study of fluency in English output of Chinese consecutive interpreting learners. *Journal of Zhejiang International Studies University, 1*, 60–68.

Yu, W.-T., & Van Heuven, V. J. (2017). Predicting judged fluency of consecutive interpreting from acoustic measures: Potential for automatic assessment and pedagogic implications. *Interpreting, 19*(1), 47–68. https://doi.org/10.1075/intp.19.1.03yu

Yuan, Y. (2016). A feature set for automated human translation quality estimation. *Foreign Language Teaching and Research, 49*(5), 776–787.

**Biographies**

Chao Han is a professor in the College of Foreign Languages and Cultures at Xiamen University, China. His research interests include testing and assessment issues in translation and interpreting, and methodological aspects of translation and interpreting studies. His recent publications have appeared in peer-reviewed journals such as *Interpreting, The Translator, Language Testing, Language Assessment Quarterly*, and *Assessment and Evaluation in Higher Education*. He is currently a member of the Advisory Board of *Interpreting: International Journal of Research and Practice in Interpreting*.

Xiaolei Lu is an assistant professor in the College of Foreign Languages and Cultures at Xiamen University, China. Her research interests include corpus processing, translation technology, and automated translation assessment. Her articles have appeared in peer-reviewed journals such as *Natural Language Engineering* and *Journal of Chinese Information Processing*. She is the co-author of *Applied Corpus Processing With Python*, a volume dedicated to corpus data processing and analysis.