Meta
Journal des traducteurs
Translators' Journal

META

Analytic rubric scoring versus comparative judgment: a comparison of two approaches to assessing spoken-language interpreting

Chao Han

Volume 66, Number 2, August 2021

URI: https://id.erudit.org/iderudit/1083182ar
DOI: https://doi.org/10.7202/1083182ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Han, C. (2021). Analytic rubric scoring versus comparative judgment: a comparison of two approaches to assessing spoken-language interpreting. *Meta*, *66*(2), 337–361. https://doi.org/10.7202/1083182ar

Article abstract

In this article, we report on an empirical study conducted to evaluate the utility of analytic rubric scoring (ARS) vis-à-vis comparative judgment (CJ) as two approaches to assessing spoken-language interpreting. The primary motivation behind the study is that the potential advantages of CJ may make it a promising alternative to ARS. When conducting CJ on interpreting, judges need to compare two renditions and decide which one is of higher quality. Such binary decisions are then modeled statistically to produce a scaled rank order of the renditions from "worst" to "best." We set up an experiment in which two groups of raters/judges of varying scoring expertise applied both CJ and ARS to assess 40 samples of English-Chinese consecutive interpreting. Our analysis of quantitative data suggests that overall ARS outperformed CJ in terms of validity, reliability, practicality and acceptability. Qualitative questionnaire data helped us obtain insights into the judges'/raters' perceived advantages and disadvantages of CJ and ARS. Based on the findings, we tried to account for CJ's underperformance vis-à-vis ARS, focusing on the specificities of interpreting assessment. We also propose potential avenues for future research to improve our understanding of interpreting assessment.

Analytic rubric scoring versus comparative judgment: a comparison of two approaches to assessing spoken-language interpreting

CHAO HAN
Xiamen University, Xiamen, China
chaohan@xmu.edu.cn

RÉSUMÉ

Dans cet article, nous rendons compte d'une étude empirique menée pour évaluer l'utilité de la notation analytique (ARS) par rapport au jugement comparatif (CJ) en tant que deux approches pour évaluer l'interprétation en langue parlée. La principale motivation derrière l'étude est que les avantages potentiels du CJ peuvent en faire une approche prometteuse par rapport à l'ARS. Lors de la conduite de CJ sur l'interprétation, les juges doivent comparer deux interprétations et décider laquelle est de meilleure qualité. Ces décisions binaires sont ensuite modélisées statistiquement pour produire un ordre de classement à l'échelle des rendus du « pire » au « meilleur ». Nous avons mis en place une expérience dans laquelle deux groupes d'évaluateurs/juges de différentes expertises de notation ont appliqué CJ et ARS pour évaluer 40 échantillons d'interprétation consécutive anglais-chinois. Notre analyse des données quantitatives suggère que l'ARS globale a surpassé CJ en matière de validité, fiabilité, praticité et acceptabilité. Les données du questionnaire qualitatif nous aident à obtenir un aperçu des avantages et des inconvénients perçus par les juges/évaluateurs de CJ et ARS. Sur la base des résultats, nous avons essayé de tenir compte de la sous-performance de CJ vis-à-vis de l'ARS, en nous concentrant sur les spécificités de l'interprétation de l'évaluation. Nous proposons également des pistes de recherche futures pour améliorer notre compréhension de l'évaluation de l'interprétation.

ABSTRACT

In this article, we report on an empirical study conducted to evaluate the utility of analytic rubric scoring (ARS) vis-à-vis comparative judgment (CJ) as two approaches to assessing spoken-language interpreting. The primary motivation behind the study is that the potential advantages of CJ may make it a promising alternative to ARS. When conducting CJ on interpreting, judges need to compare two renditions and decide which one is of higher quality. Such binary decisions are then modeled statistically to produce a scaled rank order of the renditions from "worst" to "best." We set up an experiment in which two groups of raters/judges of varying scoring expertise applied both CJ and ARS to assess 40 samples of English-Chinese consecutive interpreting. Our analysis of quantitative data suggests that overall ARS outperformed CJ in terms of validity, reliability, practicality and acceptability. Qualitative questionnaire data helped us obtain insights into the judges'/raters' perceived advantages and disadvantages of CJ and ARS. Based on the findings, we tried to account for CJ's underperformance vis-à-vis ARS, focusing on the specificities of interpreting assessment. We also propose potential avenues for future research to improve our understanding of interpreting assessment.

RESUMEN

Este artículo expone un estudio empírico acerca de la utilidad de la notación analítica (ARS) con relación al juicio comparativo (CJ) como enfoques para evaluat la interpretación en lengua hablada. Se justifica este estudio por las ventajas potenciales del CJ

comparado con ARS. Al llevar a cabo un CJ en interpretación, los jueces deben comparar dos interpretaciones y decidir cuál es mejor. Luego se modelizan estadísticamente estas decisiones binarias para generar un orden de clasificación de las producciones de la «peor» a la «mejor». Hemos realizado un experimento en el cual dos grupos de evaluadores/jueces expertos en distintas formas de notación emplearon CJ y ARS para evaluar 40 interpretaciones consecutivas del inglés al chino. El análisis de los datos cuantitativos sugiere que ARS superó globalmente CJ en cuanto a validez, confiabilidad, practicidad y aceptabilidadmatière de validité, fiabilité, praticité et acceptabilité. Los datos del cuestionario cualitativo contribuyen a observar las ventajas e inconvenientes percibidos por los evaluadores/jueces de CJ y ARS. A partir de los resultados, hemos intentado considerar el menor despeño de CJ con relación a ARS, concentrándonos en las especificidades de la interpretation de la evaluación. Igualmente proponemos pistas de investigación para mejorar nuestra comprensión de la interpretación de la evaluación.

MOTS-CLÉS/KEYWORDS/PALABRAS CLAVE

notation des rubriques analytiques, jugement comparatif, interprétation de la langue parlée, évaluation par un évaluateur, évaluation de la qualité en interprétation
analytic rubric scoring, comparative judgment, spoken-language interpreting, rater-mediated assessment, interpreting quality assessment
notación de los rubros analíticos, juicio comparativo, interpretación de la lengua hablada, evaluación mediante un evaluador, evaluación de la calidad en interpretación

1. Introduction

Assessing interpreting quality is not an easy task. Typically, assessors need to multitask: listening to target-language renditions, reading source-language input, evaluating the goodness of the fit between source- and target-language materials on multiple quality dimensions, and providing scores and making a final verdict (see Gile 1995; Han 2015). As such, the field of Interpreting Studies has witnessed a diverse array of methods being trialed and used to assess spoken-language interpreting. The primary goal of such exploration is to help raters obtain valid and reliable scores while still achieving scoring efficiency. One method, labeled atomistic analysis, is based on a rigorous examination of points of content in an interpretation and/or its (para)linguistic features, as demonstrated by researchers conducting proposition-based analysis, error analysis and (dis)fluency analysis (for example Mead 2005; Han, Chen, *et al*. 2020). Another method is based on checklists which consist of an inventory of quality criteria, so that assessors can provide quantitative ratings based on a Likert-type scales and/or qualitative comments (see for example Hartley, Mason, *et al*. 2003; Lee 2015).

Recently, rubrics-referenced rating scales, or rubric scoring, have been increasingly used to assess spoken-language interpreting (for a review, see Han 2018b). Essentially, rubric descriptors are developed to capture typical features at different levels of a performance continuum. The use of analytic rubric scoring (ARS) has been gaining traction in interpreter education (Lee 2008), professional certification (Liu 2013) and interpretation research (Han 2018b). Preliminary evidence indicates that ARS is a valid and reliable approach to assessing interpreting (Liu 2013; Han 2015; 2017; Lee 2015), although it has several potential downsides, including difficulties of generating accurate descriptors, resource-intensiveness of rater training, and rater effects (for example rater severity), to mention but a few.

A lesser-known approach is comparative judgment (CJ), which was first trialed by Wu (2010) and most recently by Han, Chen, *et al.* (2019) to assess spoken-language interpreting. Briefly, the CJ method, rooted in psychophysical analysis (Thurstone 1927) and applied to evaluate students' performance in educational assessment (for example Pollitt 2012; McMahon and Jones 2015), requires judges to compare two like objects (in our case, two interpreted renditions) and make a binary decision about their relative qualities (that is deciding which rendition is of higher quality than the other). In CJ, interpreting quality is a global construct perceived by individual judges. The binary outcomes from repeated comparisons between different pairs of renditions are then fitted to a statistical model, yielding standardized estimates (in logits) for the quality of each rendition. These estimates can be used to locate each rendition along a continuum of perceived quality, thus creating a scaled rank order of all renditions from "worst" to "best" quality. Reportedly, the CJ method has a number of advantages over ARS (Pollitt 2012; Jones and Wheadon 2015; Steedle and Ferrara 2016), including, for instance, that there is no need to specify the construct to be assessed, to conduct extensive rater training, nor to correct for rater severity (which can be minimized by CJ).

Although both Wu (2010) and Han, Chen, *et al.* (2019) reported positive results for CJ in interpreting assessment, there has been no systematic evaluation of CJ's utility vis-à-vis that of such prevalent methods as ARS. Given the potential advantages of CJ described in the literature, and given the prospect of CJ as a viable alternative to ARS, it would be of interest to compare the reliability, validity and practicality of CJ versus ARS in interpreting assessment.

2. Literature review

This section first provides a review of ARS, highlighting its potential limitations. It then segues into an overview of CJ, expounding its underlying rationales and assumptions and historical developments. It also synthesizes the claimed advantages of CJ, each of which represents a response to the limitations of ARS. The section finishes by describing several studies in which CJ is applied to assess spoken-language interpreting.

2.1. Analytic rubric scoring

The effectiveness of ARS rests on the assumption that raters are able to compare a given performance against external and theoretical standards (often manifested by rubric descriptors), and to assign numeric scores accurately and consistently to multiple performance traits/dimensions. As such, ARS is intrinsically a form of *absolute judgment* (Laming 2004; Tarricone and Newhouse 2016). The growing espousal of ARS among educators and researchers is a testimony to its usefulness in interpreting assessment (Lee 2008; Liu 2013; Han 2018b).

However, ARS is by no means perfect. There are a few problems. First, the development and validation of rubric descriptors is onerous and challenging, especially for high-stake assessments. Second, given that rubrics are predetermined and are necessarily a partial representation of a given phenomenon, there is a concern that the scope of rubric descriptors may be narrow and constrictive, and thus they may

not do justice to inherently complex and nebulous constructs such as interpreting quality. Third, there is no guarantee that the use of ARS automatically leads to high-quality assessment. Raters may construe descriptors differently, emphasize different aspects of quality and, in the worst case scenario, even ignore descriptors altogether, relying on their own internalized standards. Because of these problems, extensive training should be conducted on a regular basis to help raters establish a common understanding of rubric descriptors and apply them consistently. Undoubtedly, such training is time-consuming and resource-intensive, which can easily take up several days, if not many hours (see Han 2018b). Despite rater training, previous research reveals undesirable rater effects in ARS-based interpreting assessment, such as rater severity (Han 2015; Wang, Napier, *et al*. 2015), possible halo effects (Wu, Liu, *et al*. 2013) and rater inaccuracy (Han 2018a). Fourth, anecdotal evidence suggests that although rubrics are intended to assist raters' decision making by providing a frame of reference, some raters may make relative judgments by comparing a given interpretation to others previously listened to, rendering external standards (for example rubrics) largely irrelevant. As a result, raters' assessments may be systematically affected by the order in which interpretations are presented (that is order/sequence effect). These potential limitations of ARS have prompted researchers to look for new scoring methods that can mitigate or even eliminate such concerns.

2.2. Comparative judgment

A potentially effective response to the limitations of ARS is comparative judgment (CJ), which is a scaling method that involves comparing one performance against another regarding a global construct. It is therefore different from ARS, which involves comparing a concrete performance against specified theoretical standards. The rationale underlying CJ is the psychophysical principle that human beings are more reliable and accurate in making *relative judgments* (that is comparing one object to another) than making absolute judgments (that is judging the value of an object in isolation) (Thurstone 1927; Laming 2004). It is claimed that CJ does not enhance assessors' ability to make judgments per se, but is able to maximize the innate human capacity for making accurate comparisons (Jones and Wheadon 2015).

CJ was initially trialed by Thurstone (1927) to scale sense impressions (for example perceived magnitude) that physical substrates such as weight, loudness and brightness have. It was then applied to psychological constructs that have no physical correlates, for instance, attitudes and social values (Thurstone 1954). Later, CJ was introduced to the field of education. One of the first published applications of CJ was Pollitt and Murray (1996) who trialed CJ to evaluate students' speaking performance. Another early application of CJ was to compare standards across equivalent forms of examination papers (Bramley, Bell, *et al*. 1998). Following that, CJ was introduced to assess a wide range of complex educational constructs, including mathematics, writing and creativity.

There are a number of requirements for operationalizing CJ. First, individual judges need to make independent comparisons. Second, it is a group of judges that undertake the CJ exercise, with each judge comparing multiple pairs of relevant objects. As such, the validity of CJ is grounded in cumulative consensus arising from repeated comparisons. That is, the iterative process of CJ helps accommodate the

potentially heterogeneous understandings of a given construct among a group of judges (Jones and Inglis 2015). Third, it is desirable that the judges involved in CJ represent a community of subject matter experts, so that a broad homogeneity across the judges' understanding of a construct could be assumed (McMahon and Jones 2015). CJ-based scaling outcomes thus assimilate and reflect the collective expertise of judges (Pollitt 2012; Jones, Swan, *et al*. 2015).

Regarding the analysis of CJ data, the binary outcomes are usually modeled via the Bradley-Terry-Luce model (Bradley and Terry 1952; Luce 1959), which can also be closely approximated by the Rasch logistic model, as demonstrated by Andrich (1978). The created scale values or estimates (in logits) represent "measures" of whatever CJ is based on (in our case, it is interpreting quality). The properties of the model, such as the self-consistency of the judges' decision making, can also be examined.

2.3. Potential advantages of comparative judgment

CJ could be a potentially attractive alternative to ARS in interpreting assessment, as its inherent features seem to assuage the above concerns associated with ARS. First, while it is difficult to create rubric descriptors that accurately and fully capture a nebulous construct, CJ does not require specific assessment criteria, as it is based on a collective understanding of the construct by a community of experts (Jones and Inglis 2015; Jones, Swan, *et al*. 2015; McMahon and Jones 2015). It is thus suited to assessing constructs "that are not readily defined and operationalized in rubrics" or "a wide variety of unpredictable responses that would be difficult to anticipate comprehensively and precisely in rubrics" (Jones and Wheadon 2015: 95). One could rightfully argue that interpreting quality is one such construct that defies thorough and precise definition (for example Grbić 2008). AIIC (1982) even refers to quality as "that elusive something which everyone recognises but no one can successfully define."

Second, CJ can be conducted without extensive training, while in ARS-based assessment, rigorous rater training is strongly advised or even mandatory (Setton and Dawrant 2016; Han 2018b). Researchers in previous applications of CJ (for example Wu 2010; Jones and Inglis 2015; Han, Chen, *et al*. 2019) only conducted minimal training such as a brief introduction to CJ. This is because CJ is based on relative judgment, relying on the judges' cumulative consensus on a given construct. It is therefore a potentially cost-effective alternative to ARS. Another incentive for promoting CJ is that rater training is often sidelined or even ignored in educational interpreting assessment (for a detailed description, see Liu, Chang, *et al*. 2008). It can thus be contended that, rather than demanding resource-intensive rater training, which is simply unaffordable and impractical for interpreting programs, the use of CJ seems to be a more sensible approach.

Third, since CJ only requires holistic, relative decisions, it may be faster and easier. For example, the original Thurstonian CJ is based on intuitive, instantaneous decision making to judge the perceived magnitude of physical attributes. Such comparisons usually only involve an immediate perception, requiring little cognitive processing effort (Bramley, Bell, *et al*. 1998; Tarricone and Newhouse 2016). In contrast, using ARS to assess interpreting is a complicated multitasking process (Han 2015; 2018b) which could saturate the short-term memory capacity of most

assessors (Gile 1995) and may be as cognitively taxing as interpreting (Wu 2010). Compared with ARS, CJ seems to be less cognitively demanding, which means less rater fatigue and more expeditious decision making.

Fourth, in ARS-based interpreting assessment, rater severity is a cause for concern, as it may distort assessment outcomes (Han 2015; Wang, Napier, *et al.* 2015). With CJ, however, such a concern is no longer pertinent. CJ experimentally removes rater effects (Andrich 1978; Pollitt 2012): no matter how severe a rater may be, the same discrimination is present across all judgments and a better performance can still be identified. The judgment process is based on the relative merit, not absolute quality, of each performance.

2.4. Assessing spoken-language interpreting: A comparative judgment approach

One of the first studies to apply CJ in spoken-language interpreting assessment was conducted by Wu (2010) who recruited 30 examiners to use CJ to assess English-to-Chinese simultaneous interpreting produced by five postgraduate students of different (known) abilities. Overall, Wu (2010) found that CJ functioned effectively, distinguishing different levels in the students' performance. More recently, CJ was also applied by Han, Chen, *et al.* (2019) to assess English-Chinese consecutive interpreting. Research results suggest that the scale of interpreting quality, constructed based on the binary CJ data, has relatively high separation reliability and that the majority of the judges behaved consistently. Both Wu's (2010) and Han, Chen, *et al.*'s (2019) studies invite further, systematic exploration of CJ as a potential alternative to ARS.

3. Research questions

We therefore conducted an experiment to evaluate how CJ would perform vis-à-vis ARS in an assessment of spoken-language interpreting. We aimed to address four research questions (RQs) concerning the validity, reliability, practicality and acceptability of CJ versus ARS.

- RQ1: How would CJ and ARS compare, in terms of (concurrent) validity?
- RQ2: How would CJ- and ARS-based measures compare, in terms of reliability?
- RQ3: To what extent is CJ more practical than ARS, regarding the amount of time needed?
- RQ4: How would judges/raters perceive the use of CJ versus ARS?

The study also attempts to compare the utility of CJ and ARS along two dimensions. First, we asked both novice and experienced judges/raters to use CJ and ARS. By doing so, we would be able to verify the assumption that the judges'/raters' expertise leads to better CJ and/or ARS outcomes. Second, we asked raters/judges to assess interpreting into both their first language (L1) and second language (L2). Previous literature on ARS shows that rubric scores were more reliable in the assessment of interpreting into their L1 than into their L2 (Han 2016; 2019).

4. Method

4.1. Participants

Two groups of judges/raters were recruited to assess interpreting. The first judge/rater group consisted of 20 undergraduate interpreting students (anonymized as J/R 01-20), with 16 of them being female and the rest being male. Their average age was 21 years old, and all had completed two mandatory courses on English-Chinese interpreting. Given their limited learning and scoring experience, they were regarded as novice judges/raters. The second judge/rater group was comprised of 20 postgraduate interpreting students (that is J/R 21-40), with 17 of them being female and the rest being male. With an average age of 25 years old, they were pursuing a postgraduate-level interpreting degree (for example *Master's in Translation and Interpreting*, *Master's in Conference Interpreting*) and all were at the end of their second year of full-time study. Considering that these students had more learning experience, constantly participated in classroom-based evaluation activities (for example self- and peer assessment) and were better acquainted with the relevant literature (thanks to a research component in the training program), it is believed that they had obtained an advanced understanding of interpreting quality. As such, they were regarded as the experienced judges/raters. All judges/raters had Mandarin Chinese as their L1 and English as their L2.

4.2. Interpreting recordings

The recordings used in the study were sourced from a larger corpus of 82 recordings collected from previous summative assessments of English-Chinese consecutive interpreting. In the assessment, 41 undergraduate students interpreted one English speech and one Chinese speech consecutively (both were about two and a half minutes or 150 seconds), and their performances were audio-recorded. Both speeches were about 400 syllables in length (the English text: 250 words; the Chinese text: 400 characters) and related to general topics, with the English speech focusing on foreign direct investment in China and the Chinese speech on changing consumption patterns in China. They were delivered at a relatively slow speed of 160 syllables/characters or 100 words per minute. As such, they were deemed appropriate for the student interpreters. For practical reasons, we carefully selected 40 recordings produced by 20 students (that is English-to-Chinese, 20 recordings; Chinese-to-English, 20 recordings) that represented different levels along the ability continuum. This deliberate selection was possible because of the availability of performance estimates (that is the achievement data) we obtained in the assessment. Regarding our selected samples, the average duration of a recording for the English-to-Chinese direction was about 150.8 seconds, while that for the other direction was 229.8 seconds.

4.3. Experimental design

The repeated-measures design we implemented consists of two rounds of performance assessment. In the first round, both groups of judges were asked to perform CJ. To minimize the order effect, we counter-balanced the interpreting directions. That means that one group of judges first conducted CJ on English-to-Chinese renditions, while the other group judged Chinese-to-English renditions. When both

groups completed the required amount of CJ, they switched to the other direction. Three weeks later, in the second round of assessment, both rater groups applied ARS to assess the same set of recordings; interpreting directions were also counter-balanced.

It is worth noting that we implemented CJ first and ARS second. Our primary reason was that, had the method of ARS been used in the first round of assessment, the raters would have been informed of and become well acquainted with the specific assessment criteria in the rating scales. The familiarity with scalar descriptors may act as a normative force that shapes the raters' understanding of interpreting quality, which would subsequently influence their judgment in CJ where no specific criteria are provided. Conducting CJ first and ARS second minimizes any "spillover" effect. Also noteworthy is that we put in place a three-week interval between the two rounds of assessment and used randomization (see below) to reduce any memory effect.

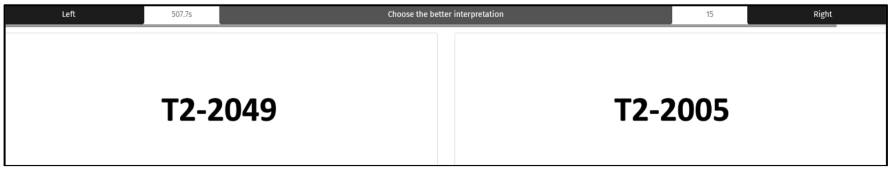
4.4. The CJ condition

To operationalize CJ, we used an online system called *No More Marking* (NMM).[1] Essentially, this system allows users to upload online scanned samples of written performance. Each judge can access the system using a unique web link. Judges are able to review a given pair of written samples presented side by side on the web browser, and select the higher quality sample by clicking a "left" or "right" decision button. The system automatically tallies the frequency of how many times a given sample beats the others, and records each judge's decision history and the amount of time taken to make a decision.

In our study, however, the samples were not written responses, but audio recordings. To make CJ possible, we first labeled the 40 recordings (for example T2-2049, meaning that the recording was produced by student #2049 in task #2). We then copied each of the 40 labels into a PDF file. It was the PDF files that were uploaded to the NMM system so that a random sequence of paired recordings could be generated for each judge (for an example, see Figure 1). Based on the NMM-generated sequence of pairings, the judges selected the paired recordings from their computer, listened to them at their own pace, and made comparative judgments. This arrangement was repeated for both interpreting directions.

Figure 1

An example of paired comparison realized through the NMM website



During the training provided to the judges, which was approximately 30 minutes long, the raters were asked to familiarize themselves with the source-language texts and they were given a brief explanation of the NMM system. We asked them to make independent judgments according to their understanding of interpreting quality.

With 20 recordings for each direction (that is n = 20), the number of possible pairwise comparisons was n×(n-1)/2 = 190. This was far too large a number for a single judge in our exploratory study. According to previous research (Pollitt 2012; McMahon and Jones 2015), a satisfactory scale separation can be achieved, based on a small number of pairings, so long as data connectivity is established among comparisons. The NMM website suggests that 10 judgments be conducted by each judge to obtain relatively stable parameter estimates. We therefore required that each rendition be judged 15 times.

4.5. The ARS condition

The two rater groups used an eight-point, rubrics-referenced analytic rating scale to assess interpreting, focusing on three major dimensions: 1) information completeness (InfoCom: to what extent original content is successfully rendered); 2) fluency of delivery (FluDel: to what extent disfluencies, such as (un)filled pauses, long silence, fillers and/or excessive repairs are present in renditions); and 3) target language quality (TLQual: to what extent target-language expressions are natural to a native English or Chinese speaker) (see Appendix). Previous studies suggest the sound psychometric properties of the three sub-scales (for example the ability to distinguish different levels of performance) (Han 2015; 2017).

It is typical of ARS to provide rigorous training to all raters so that they could achieve a consistent understanding of scalar descriptors. However, in this study, we deliberately provided very brief training, of about 30 minutes, to be on a par with the CJ condition: we first asked the raters to familiarize themselves with the source-language texts (about 10 minutes), then we introduced the rating scale (for example its format, quality criteria, scalar descriptors), and finally explained how the scale needed to be used (for example giving an integer score of one to eight). By doing so, we wanted to make sure that, in both CJ and ARS conditions, rater training was comparable. If the utility of ARS was approximate to or even outperformed that of CJ, we would have good reasons to believe that CJ may not be a promising alternative to ARS.

All raters assessed interpreting independently. To make it comparable to the CJ condition, each rater assessed 15 randomly selected recordings for each direction. The order of recordings to be assessed was also randomized for each rater. In both scoring conditions (CJ and ARS), to evaluate the interpreting quality, the judges/raters listened to the recordings while checking against the written source-language text.

4.6. Post-hoc questionnaires

Once the judges completed the first round of assessment (that is the CJ condition), they were asked to comment on the use of CJ by answering a short questionnaire (Questionnaire A).[2] In the wake of the second round of assessments (that is the ARS condition), they filled out another short questionnaire (Questionnaire B) that tapped into the raters' perceptions of ARS.[3] The content of the questionnaires pertained to: 1) to what extent the judges/raters were confident, based on a five-point Likert scale, in using CJ/ARS to assess interpreting in both directions (Items 1 & 2); 2) what were

the advantages and disadvantages of CJ/ARS (Items 3 & 4); and 3) what quality criteria did the judges rely on when making CJ decisions (Item 5, only in Questionnaire A). The purpose of incorporating a survey component into the study was to obtain a primarily qualitative understanding of how CJ and ARS were received by the judges/raters.

4.7. Data analysis

The raw data from both scoring conditions were modeled statistically to construct a scaled rank order of renditions from "best" to "worst." More specifically, the Bradley-Terry-Luce model was used to analyze the CJ data, which was automatically conducted by the NMM system. The ARS data was analyzed by many-facet Rasch measurement through the FACETS 3.71.0 program (Linacre 2021).[4] The parameter estimates and statistics from the above modeling were used in some of the following analyses.

To examine the concurrent validity of CJ- and ARS-based measurements, we followed previous researchers (Jones and Wheadon 2015; Steedle and Ferrara 2016) to correlate CJ and ARS measures with the actual achievement data (that is the marks we obtained from the final examination based on holistic scoring). Pearson's correlation coefficients were computed as a proxy of what is known as validity coefficients.

To understand the reliability of both methods, we generated two types of statistical evidence. One type of evidence is the replicability of CJ and ARS measures across different groups of judges/raters (see Jones and Wheadon 2015). We correlated the CJ and ARS measures produced by the novice judges/raters with those by the experienced judges/raters (that is Pearson's correlation coefficients). The other type of evidence is the judge's/rater's fit statistics for the CJ and ARS conditions (see Steedle and Ferrara 2016). This type of psychometric indices can serve as an indicator of a judge's/rater's internal self-consistency.

In terms of the practicality of CJ and ARS, our starting point was to focus on the amount of time required by CJ versus ARS. Given the same scoring workload, if one method needed more time, it would then be less cost-effective. The NMM system records the time taken for each CJ decision. We also asked the raters to estimate how long it typically took to assess a recording under the ARS condition. Invariably, the raters reported that, except for the time spent listening to the recordings (that is about two and a half minutes), assigning scores took less than one minute.

Finally, to explore the acceptability of CJ and ARS, we relied primarily on the survey data. We conducted a three-way mixed ANOVA with two within-subjects factors (that is assessment method, directionality) and one between-groups factor (that is rater type) on the quantitative confidence rating data[5], so as to explore whether the raters' confidence level differed between the assessment methods, interpreting directions and rater groups, as well as whether there was any interaction between the within-subjects factors and the between-subjects factor on rating confidence. We also content-analyzed the qualitative data iteratively to identify emergent themes from the judges'/raters' comments, and quantified the qualitative data by using frequency count statistics.

5. Results

5.1. Concurrent validity

Pearson's correlation coefficients (that is validity coefficients) calculated between the CJ data and the achievement data as well as between the ARS data and the achievement data are presented in Table 1.

TABLE 1
Correlation between ARS/CJ data and achievement data

Judge/Rater type	Correlation between ARS and achievement data		Correlation between CJ and achievement data	
	E-C	C-E	E-C	C-E
Novice	0.89**	0.80**	0.81**	0.69**
Experienced	0.87**	0.87**	0.87**	0.79**

Notes: ** $\rho < 0.01$; E-C = English-to-Chinese, C-E = Chinese-to-English.

Overall, averaged across rater groups and interpreting directions, Pearson's *r* for the CJ condition (that is (0.81+0.69+0.87+0.79)/4 = 0.79) was lower than that of the ARS condition (that is (0.89+0.89+0.87+0.87)/4 = 0.86), suggesting that the CJ measures have slightly lower concurrent validity. In addition, validity coefficients averaged across the two directions were consistently lower in the CJ than in the ARS condition, irrespective of the rater types (that is novice: (0.81+0.69)/2 = 0.75 < (0.89+0.80)/2 = 0.85; experienced: (0.87+0.79)/2 = 0.83 < (0.87+0.87)/2 = 0.87). This pattern seemed to hold true when correlation coefficients were averaged across rater types (that is English-to-Chinese: (0.81+0.87)/2 = 0.84 < (0.89+0.87)/2 = 0.88; Chinese-to-English: (0.69+0.79)/2 = 0.74 < (0.80+0.87)/2 = 0.84). Finally, examining each individual Pearson's *r* in Table 1, we find that in each case the validity coefficient was larger in the ARS than in the CJ condition, except only one case in which the experienced judges/raters assessed the English-to-Chinese interpreting.

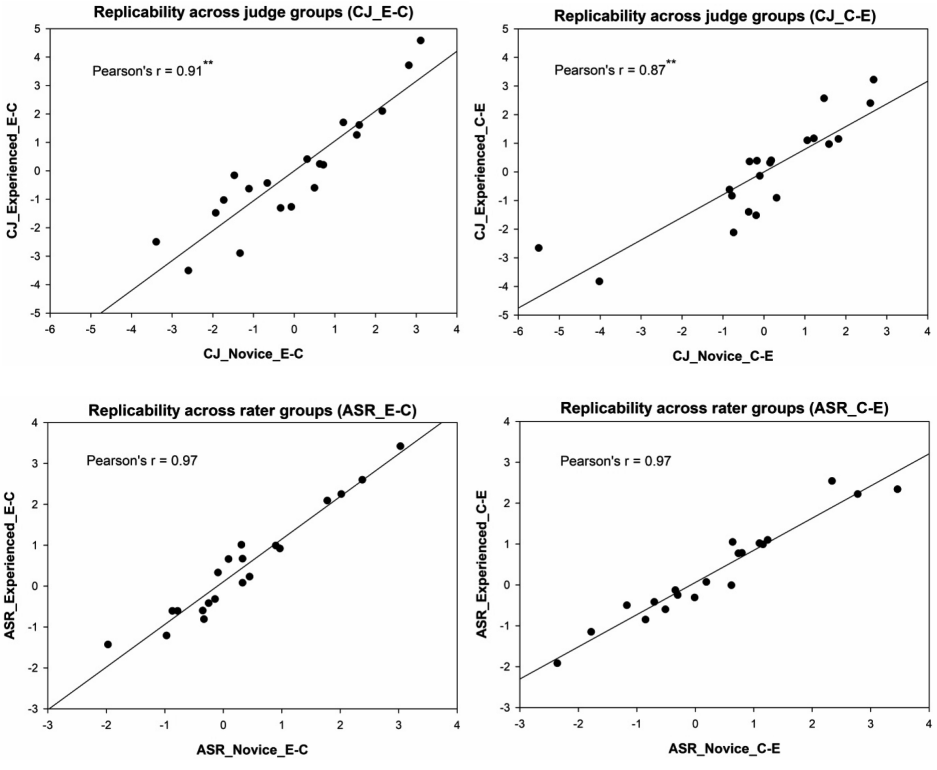
When comparing each assessment condition individually, we also found that in almost all cases the experienced judges/raters outperformed their novice counterparts (that is higher validity coefficients), except for the English-to-Chinese interpreting in the ARS condition (experienced: 0.87, novice: 0.89). Similarly, the validity of the English-to-Chinese interpreting assessments tended to be higher than that of the opposite direction in all cases but one (that is when the experienced raters used ARS).

5.2. Reliability/Replicability

We first examined the reliability of CJ and ARS regarding the replicability of CJ and ARS results across the rater groups. We correlated the CJ data (or the ARS data) produced by the novice and experienced judge/rater groups. As displayed in Figure 2, for both directions, the ARS measures were more replicable than the CJ measures (that is English-to-Chinese: 0.97 > 0.91; Chinese-to-English: 0.97 > 0.87).

Figure 2

Correlation of CJ and ARS results between rater groups



We then investigated the rater's internal consistency, based on fit statistics. More specifically, the rater's mean-squared infit statistic was used to gauge whether a rater had behaved consistently compared to the other raters in the same group. Given the exploratory nature of the study, we chose an infit statistic between 0.5 and 1.5 (that is 0.5 < infit < 1.5) to determine an acceptable level for a rater's internal consistency (see Linacre 2002). Raters whose infit values are equal to or under 0.5 are overly predictable (also known as an *overfit*); and raters whose infit values are equal to or larger than 1.5 are erratic in their decision making (that is a *misfit* or *underfit*). Overall, misfit causes more disturbances to measurement than overfit.

In Figures 3 and 4, we plotted the infit values for both rater groups in the CJ and the ARS conditions. As can be seen in the figures, dashed lines were plotted at 1.5 and 0.5. Raters with their infit values falling between the dashed lines were relatively self-consistent in their decision making. A misfit would be above the upper dashed line, whereas an overfit would be below the lower dashed line. In the figures, we used the symbol ▲ and exact fit values to display an aberrant rater. An inspection of the figures reveals that overall the number of overfits was slightly greater than that of misfits. However, it seems that there was no substantial difference between the CJ and the ARS conditions: five misfits and 10 overfits in total for the CJ condition; six misfits and seven overfits for the ARS condition.

FIGURE 3

Display of infit statistics for both rater groups in the CJ condition

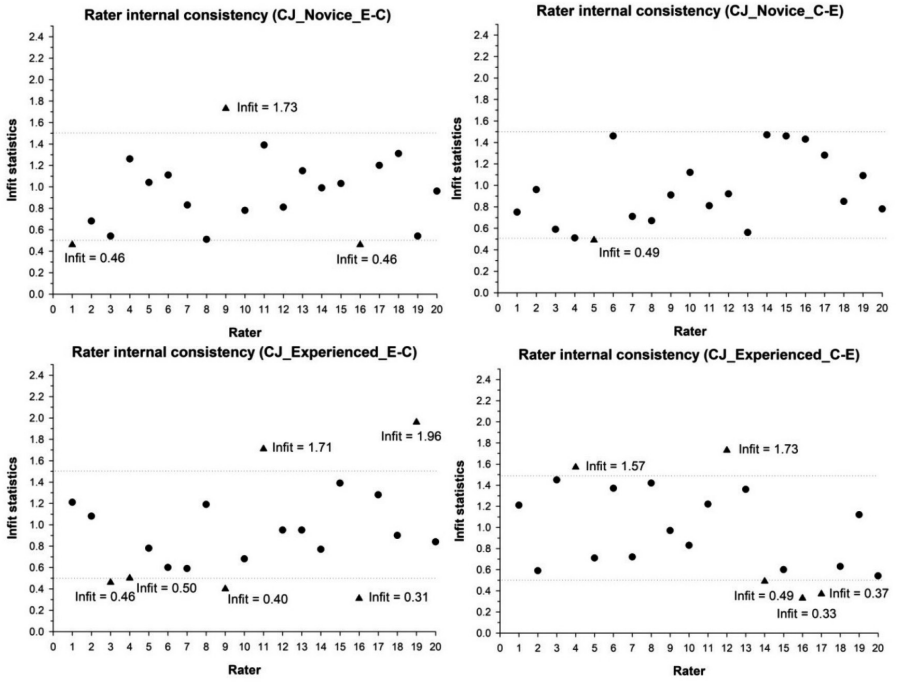
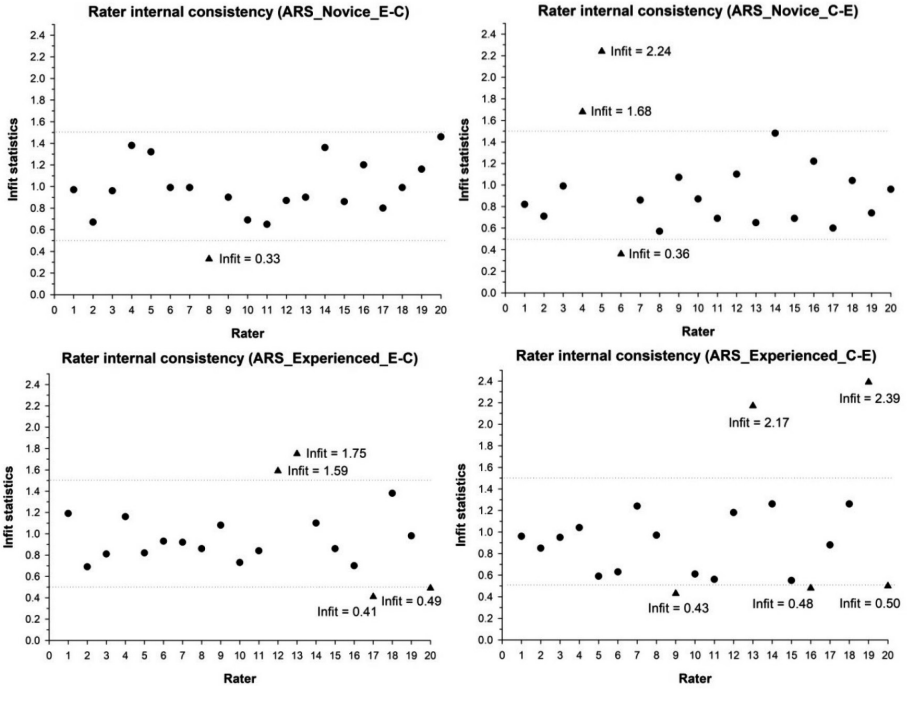


FIGURE 4

Display of infit statistics for both rater groups in the ARS condition



5.3. Practicality of CJ and ARS

As the NMM system recorded the exact time taken to make CJ decisions, we were able to calculate the average time spent by each rater group in each direction, as is shown in Table 2. According to the table, for each direction, the experienced raters seemed to work faster than their novice counterparts.

Table 2

The average amount of time spent on making 15 paired comparisons

Rater Group	Interpreting Direction	
	E-C	C-E
Novice Raters	5646.86s (94 min)	7140.45s (119 min)
Experienced Raters	4976s (83 min)	6188.51s (103 min)

Notes: s = second, min = minute; E = English, C = Chinese

When it comes to the ARS condition, we were only able to provide an estimate of the amount of time taken for assessing 15 recordings. Given that the raters reported that they had spent approximately one minute to assign ratings after listening to the recordings, we estimated that the average time for a rater to assess 15 recordings was 3162 seconds (or 53 minutes) for the English-to-Chinese direction (that is 15 recordings × 150.8 seconds + 15 recordings × 60 seconds = 3162 seconds) and 4346.25 seconds (or 72 minutes) for the opposite direction (that is 15 recordings × 229.75 seconds + 15 recordings × 60 seconds = 4346.25 seconds). Two types of time were included in the calculation: 1) the total amount of time the raters used to listen to the recordings, and 2) the total amount of time spent on assigning ratings.

From another perspective, in the CJ condition, an average judge spent about 354 seconds (or 5.9 minutes) to make a single paired comparison in the English-to-Chinese direction, and about 444 seconds (or 7.4 minutes) in the Chinese-to-English direction. By contrast, in the ARS condition, an average rater used about 210 seconds (or 3.5 minutes) and 290 seconds (or 4.8 minutes) to assess a single recording.

5.4. Acceptability of CJ and ARS

The mixed ANOVA analysis indicates that both the assessment method and interpreting directionality had a statistically significant effect on the raters' confidence level: F (1, 39) = 9.02, $\rho < 0.01$ and F (1, 39) = 13.98, $\rho < 0.01$, respectively. Specifically, the raters were more confident using the ARS method (M = 3.98, SD = 0.40) than the CJ method (M = 3.73, SD = 0.74); they were also more confident assessing the English-to-Chinese direction (M = 4.01, SD = 0.53) than the opposite direction (M = 3.69, SD = 0.61). However, although on average the novice raters (M = 3.79, SD = 0.59) seemed to have reported a lower level of confidence than their experienced counterparts (M = 3.91, SD = 0.53), we did not find any statistically significant difference between the rater groups, F (1, 19) = 1.10, $\rho = 0.30$. In addition, regarding the interaction effect, only one statistically significant interaction effect was found between the assessment method and the interpreting direction, F (1, 19) = 4.22, $\rho < 0.05$.

Based on the qualitative data from the questionnaires, we identified possible factors that could account for some of the statistically significant differences. For example, the greater confidence level observed for the English-to-Chinese direction could be explained by language familiarity. The judges/raters predominantly associated higher confidence levels when assessing interpreting into their L1, which can be illustrated by the quote from J/R 16:

> Chinese is my mother tongue, whereas English is my L2. My Chinese proficiency is, with no doubt, higher than my English proficiency in every aspect… I was more confident when assessing English-to-Chinese interpretation, because I was capable of evaluating the quality of Chinese expression… in contrast, with English being my L2, I have much room for improvement as an assessor. Particularly, I was not sure whether the interpreters had expressed the original message idiomatically and naturally… (from J/R 16)

In addition, the lower confidence level observed in the CJ condition could be partly attributed to the lack of common assessment criteria. Some judges expressed that they were unable to rely on a coherent set of criteria and, as a result, may focus on different sets of criteria from one judgment to another, demonstrated by the following quote:

> When I was conducting paired comparison, I first evaluated whether an interpretation was meaningful and comprehensible; I then examined whether there were multiple unnecessary pauses; finally, I decided on whether the delivery was natural and confident. With these being said, there were no fixed criteria for me, really. My decision was based primarily on my intuition; it was not objective… (from J/R 10)

Furthermore, the content analysis casts light on the judges'/raters' perceived advantages and disadvantages of CJ and ARS. We want to showcase the most frequently mentioned themes, and also highlight a few thought-provoking revelations. Regarding CJ, we identified five themes related to its advantages (labeled as CJ-A) and seven themes to its disadvantages (CJ-D). We presented these themes, together with their respective frequencies, in Table 3. More specifically, the judges offered some insightful comments (illustrated by the following quotes concerning CJ-A1, CJ-A3, CJ-D1 and CJ-D6) that help enrich our understanding of CJ.

> CJ-A1: In many cases, I was able to make instantaneous decisions. Plus, it seems that human beings have long been good at making pairwise comparisons, an innate ability indeed (from J/R 39).

> CJ-A3: In hindsight, I think that comparing two recordings was a natural and unconscious process of holistic evaluation. That is, when two interpreters were both good at rendering original messages faithfully, I tended to heed other dimensions of the renditions, thus turning the evaluation into a multi-criterial analysis (from J/R 34).

> CJ-D1: Sometimes, both renditions grouped in a pairing were more or less of the same quality, be it good or bad. I was therefore unable to make accurate judgment (from J/R 03).

> CJ-D6: When I was listening to the second recording in a pairing, my overall impression of the first recording tended to fade from my memory. As a result, I may not be able to judge their quality objectively (from J/R 34).

Table 3
Description of identified themes and their frequencies for CJ

CJ	Code	Description of Themes	Frequency
Advantage (A)	CJ-A1	• CJ embodied relative judgment, which was fast and accurate.	18
	CJ-A2	• CJ helped identify strengths and weaknesses of a rendition.	13
	CJ-A3	• CJ represented a multi-criterial, holistic approach to interpreting evaluation.	5
	CJ-A4	• CJ cancelled out rater bias.	3
	CJ-A5	• CJ produced reliable outcomes thanks to a series of repeated comparisons.	2
Disadvantage (D)	CJ-D1	• It was difficult to make CJ decisions when two renditions were of similar quality.	10
	CJ-D2	• Different assessment criteria were used by different judges and across comparisons.	9
	CJ-D3	• CJ was time-consuming because I needed to listen to recordings multiple times.	7
	CJ-D4	• The relativity of CJ did not say much about the absolute quality of a rendition.	5
	CJ-D5	• It was possible to encounter the same rendition repeatedly in CJ.	3
	CJ-D6	• Judges' memory capacity may play a role in CJ because the first rendition needed to be remembered for later comparison.	2
	CJ-D7	• The quality of the first rendition listened to affected my judgment on the second rendition.	2

Similarly, we found five themes associated with the advantages of ARS (labeled ARS-A), and nine themes with its disadvantages (ARS-D). Table 4 displays a description of these themes and their respective frequency. We also provided direct quotes to illustrate a number of interesting themes, below:

> ARS-A1: The advantage of rubric scoring was that it provided assessors with a common set of criteria as a reference, which is a prerequisite for the consistency and objectivity of assessment results (from J/R 20)

> ARS-A4: ARS helped me to gain a detailed understanding of interpreting performances for each scale band. I was able to understand what a good rendition or a bad rendition looked like. I was also able to evaluate what levels I myself could achieve, if I interpreted for the speeches, and what I needed to do to close the gaps between my performance and an excellent rendition described in the scale (from J/R 21).

> ARS-D6: I felt that the three assessment criteria would influence one another. For example, when I decided that InfoCom for a given interpretation was not good, I would unconsciously choose the same band for the other two performance dimensions (J/R 15).

> ARS-D9: I tended to evaluate a rendition by comparing it to a previous rendition. Suppose that there were two recordings: A2 and A10, and that A10 was actually better

than A2. However, A2 and A10 may be awarded with the same score, based on ARS. This is because the recording assessed prior to A2 was very bad, which made A2 sound pretty good. Meantime, the recording presented before A10 was really a good one, A10 was then judged as an average performance (from J/R 25).

TABLE 4
Description of identified themes and their frequencies for ARS

ARS	Code	Description of Themes	Frequency
Advantage (A)	ARS-A1	• ARS provided a consistent and objective frame of reference for evaluation.	18
	ARS-A2	• The analytic nature of ARS was useful.	11
	ARS-A3	• ARS quantified interpreting quality directly, so that differences were easy to understand.	8
	ARS-A4	• ARS was a criterion-referenced approach to evaluation.	7
	ARS-A5	• The assessment criteria were comprehensive.	6
Disadvantage (D)	ARS-D1	• The scalar descriptors (especially for FluDel and TLQual) needed to be fine-tuned.	10
	ARS-D2	• The analytic scale needed to incorporate more quality criteria.	9
	ARS-D3	• More performance levels could be included to increase measurement precision.	9
	ARS-D4	• When using ARS, raters were subjective and may construe assessment criteria differently.	6
	ARS-D5	• To use the scale properly, one needed to have sufficient scoring expertise.	4
	ARS-D6	• The three assessment criteria may interfere with one another (that is halo effect).	4
	ARS-D7	• There were no concrete exemplars anchored to each performance level.	4
	ARS-D8	• There were no weighting schemes applied to the current rating scale.	2
	ARS-D9	• The quality of a previous rendition affected my evaluation of the next one (that is order effect).	2

Finally, we content-analyzed the judges' responses to Item 5 in Questionnaire A (that is what aspects of interpreting did the judges rely on to make CJ decisions). In total, we identified 16 features that could be categorized to three general quality dimensions: content, delivery and language use. Each judge reported an average of four features on which they based their CJ decisions. Table 5 displays the frequency of each feature mentioned by the judges.

Table 5
Reported features the judges relied on in CJ

General Dimension	Features Concerning Quality	Frequency
Content	• Information completeness	27
	• Information accuracy	16
Delivery	• Fluency & fluidity	28
	• Confident and pleasant voice	18
	• Native-like pronunciation	16
	• Pauses	9
	• Fillers	9
	• Excessive self-pairs	5
	• Enunciation	4
	• Proper voice volume	3
	• Hesitation	1
	• Excessive repetitions	1
Language Use	• Idiomatic language use	12
	• Lexical choice	8
	• Grammatical correctness	7
	• Logic & coherence	4

6. Discussion

6.1. Validity and reliability

Our quantitative analysis indicates a relatively high level of concurrent validity of CJ, as its average validity coefficient was 0.79; and our qualitative analysis provides some initial evidence of construct validity as well, since the quality criteria used in CJ (see Table 5) largely corresponded to those in previous assessment practices (see for example Han 2015; Lee 2015; Setton and Dawrant 2016). In addition, we found empirical evidence to the claim that CJ rests on the expertise of judges, as the experienced judges invariably outperformed their novice counterparts when implementing CJ (that is higher validity coefficients). Moreover, we found that the validity coefficients associated with the CJ measures tended to be lower than those of the ARS measures, indicating its lesser concurrent validity. This finding, however, contradicts those reported in previous studies (see Jones and Wheadon 2015; Steedle and Ferrara 2016), in which CJ measures correlated more closely with criterion measures than rubric scores.

Regarding reliability, the CJ measures were largely replicable (in terms of rank-ordering) across the judge groups, as Pearson's r was above 0.85 for both interpreting directions. This is consistent with what is reported in Jones and Wheadon (2015) (that is Pearson's $r > 0.8$). However, in comparison, we found that CJ fared no better than ARS: despite the fact that the fit analysis reveals no apparent difference between CJ and ARS regarding raters' internal consistency, the CJ measures were less replicable

across the judge groups than the ARS measures for both directions. Although the above evidence suggests that overall CJ is a valid and reliable method to assess interpreting, it failed to outperform ARS on any statistical criteria we examined.

A number of factors may account for CJ's underperformance in assessing interpreting. Chief among them is, arguably, the difference between the vintage Thurstonian CJ and the CJ practiced in the study. The traditional method of CJ requires judges to compare magnitude of simple physical properties such as weight, loudness and brightness, which is fast and barely requires much cognitive processing effort (see Thurstone 1927). However, the CJ operationalized in our study requires judgment of constructed spoken responses that differ substantively from those involved in traditional CJ. The object of comparison in our study (that is the three-minute interpreted renditions) is much lengthier and more sophisticated than sense impressions. As such, CJ in our study is more likely to require longer time and more cognitive effort, which may render relative judgment less efficient and obviate its potential advantages.

Second, although previous research involving constructed responses (for example essay writing) has reported better validity of CJ than ARS (see Steedle and Ferrara 2016), spoken-language interpreting has two inherent characteristics that would tax the judges' cognitive resources more greatly than previous written/spoken responses. One characteristic pertains to its aural modality. Although both written samples (for example essays) and interpreted renditions are typically lengthy and complicated constructed responses, the former is in a written mode (that is directly observable, stable, eternal) and the latter a spoken mode (that is dynamic, ephemeral, transient). The modality difference presents a cognitive challenge of varying degrees to judges in CJ. When judges read and compare two essays displayed side by side, they are able to switch back and forth between the two written samples, identifying and evaluating similarities and differences. With regard to interpreting, however, judges are unable to compare the quality of two renditions directly because interpreting is ephemeral (Gile 1995; Wu 2010). To make a comparison, judges need to construct and store in working memory mental representations of each rendition's quality. This extra load on working memory may cause greater cognitive complexity for judges in interpreting assessment than essay evaluation. The other characteristic of interpreting quality assessment relates to an additional process of evaluative comparison between target-language output and source-language input in terms of informational correspondence (or fidelity). This type of additional exertion of cognitive effort to ascertain fidelity is not required when evaluating monolingual writing/speaking performance. We therefore believe that judging interpreting quality not only differs substantially from the original Thurstonian CJ, but also deviates qualitatively from judging writing/speaking.

Third, the type of rubric scoring we implemented in the study may be another reason. In the literature, three types of rubric scoring have been documented (see Han 2018b): 1) descriptor-based analytic scoring used in the current study (that is ARS), 2) descriptor-based holistic scoring, and 3) impressionistic overall scoring based on short descriptors like "highest quality" and "lowest quality." Using ARS in our study, each rater was able to generate three ratings/data points (that is InfoCom, FluDel and TLQual) for each recording. And in our Rasch analysis, an average of 45 ratings (a range of 33 to 54 ratings) was available to calibrate an overall measure

of quality for each rendition. With more data points per rendition, the ARS measures seemed to be more accurate and stable. By contrast, in CJ each rendition was judged 15 times (that is 15 judgmental scores). Fewer data points per rendition were generated for statistical estimation, possibly resulting in less accurate CJ measures. It appears that the inherent advantage of ARS, being able to produce multiple ratings for each rendition, makes its measures more accurate, especially when each rendition is evaluated multiple times by multiple raters. To level the playing field, future researchers could compare results from impressionistic overall scoring (IOS) and CJ. As a variant of rubric scoring, IOS is more comparable to CJ than ARS in at least two important ways: 1) both IOS and CJ involve holistic evaluation based on a global construct of interpreting quality; and 2) both generate the same amount of data points for each rendition. In addition, IOS epitomizes absolute judgment as described in the literature (that is assigning a single absolute score to a given response), which is in direct comparison with relative judgment, exemplified by CJ.

Fourth, the judges'/raters' level of expertise could be another factor. Although our study distinguishes two levels of judging/scoring expertise, both judge/rater groups were comprised of student interpreters (that is non-experts) whose understanding of quality is still being shaped, thus less stable and coherent than true experts. When using CJ in which no specified assessment criteria are provided, individual judges in our study may find it difficult to use a consistent set of quality criteria across multiple instances of paired comparisons (see CJ-D2 in Table 4). When using ARS, however, the raters could rely on the transparent and detailed scalar descriptors as a stable frame of reference (see ARS-A1 in Table 5). It would seem that overall the normative force of ARS outplayed CJ's claimed strength of cumulative consensus. This result may not hold true when expert judges/raters were used, who have already formed a solid, stable and well-rounded understanding of quality. Ideally, such expert judges/raters are exemplified by those who have obtained a balanced mix of relevant academic qualifications, interpreting practice, teaching experience and scoring expertise. However, recruiting even a small group of such judges/raters is difficult in interpreting research (Setton and Dawrant 2016; Han 2018b).

Finally, we have to keep in mind other potentially relevant factors, including 1) the lack of optimal pairing of renditions in the CJ condition (see CJ-D1 in Table 3), and 2) the use of a single criterion measure (that is the achievement data), which could be extended to multiple criterion measures.

6.2. Practicality

The evidence from our study indicates that CJ was more time-consuming than ARS for both directions, collectively or individually. This finding challenges the previous assertion that CJ is more efficient for making a single judgment than providing a rubric score (Steedle and Ferrara 2016). It should also be noted that we provided minimal rater training in both CJ and ARS conditions, which further supports that CJ seems to be less efficient than ARS in interpreting assessment. However, the expertise of the judges/raters may account for the inconsistent finding. In Steedle and Ferrara's (2016) study, teachers were trained as CJ judges, whereas we only recruited interpreting students. In our study, on average, the student judges used twice as much time as the duration of an original recording when making a CJ decision, implying

that the students listened to paired recordings in their entirety. Had expert judges been recruited to conduct CJ, they would have employed such time-saving strategies as selective listening and spot-checking (see Jones and Inglis 2015).

6.3. Acceptability

Our analysis of the quantitative questionnaire data suggests that overall the judges/raters were more confident in using ARS than CJ. This result seems to corroborate the above findings. In addition, the judges/raters reported higher confidence in decision making for the English-to-Chinese direction than the other direction. Our analysis of the qualitative questionnaire data offers insights into how ARS and CJ were perceived by the judges/raters. We highlight a number of interesting observations, some of which could lead to new hypotheses to be further tested.

First, the judges/raters correctly pointed out that the CJ operationalized in the current study is essentially relative judgment (that is CJ-A1), which stands in contrast to ARS that rests on external standards as a frame of reference (that is ARS-A1, ARS-A4).

Second, the way two renditions are paired could affect the cognitive complexity of CJ (that is CJ-D1). We may further hypothesize that: 1) the more similar the quality of two renditions is, the more cognitively-taxing a paired comparison tends to be, and vice versa; 2) the more cognitively complex a comparison is, the less accurate judgment tends to be. Such hypotheses could be tested in future research.

A related topic has to do with the role of working memory in CJ, as judges need to remember key features of the first rendition and compare them with those of the second (that is CJ-D6). A further hypothesis would be that judges with a higher working memory capacity outperform those with a lower capacity when using CJ to assess interpreting.

Third, regarding ARS, the raters observed that the order in which renditions were assessed could affect raters' scoring (that is ARS-D9). The serial order effect has, however, received little attention in interpreting assessment. Interested researchers can look into this issue in depth.

Fourth, also regarding ARS, the halo effect may be present, since some raters reported that it was difficult to distinguish the assessment criteria (that is ARS-D6). In a previous study on scale-based interpreting assessment, Wu, Liu, *et al.* (2013) did not observe the halo effect. However, given that the analytic scale in Wu, Liu, *et al.* (2013) had only two dimensions, that is, content and delivery, the halo effect may be less likely to be observed. When multiple assessment criteria (for example $n \geq 3$) are included in ARS, there is an increasing possibility that raters will find it difficult to reliably tell them apart, hence the halo effect. Such a hypothesis could be investigated further.

Fifth, we argue that a number of reported disadvantages of ARS (that is ARS-D3, D4, D5 and D7) are due to the lack of rigorous rater training. Had such training been provided, we would have been able to enhance the raters' understanding of scalar descriptors, scale format and exemplars for each performance level, which may help further improve rater consistency.

Finally, despite the respective benefits and limitations associated with ARS and CJ, their use may ultimately depend on the purpose of assessment. In interpreter

training, if the goal of assessment is to rank-order students or to identify better performers among a group of students, CJ seems to be better suited to such a purpose. However, if diagnostic information is valued, ARS can generate a score profile that indicates the strengths and weaknesses of each student, while CJ, as implemented in the current study, is incapable of producing such information.

7. Conclusion

We set out to evaluate the utility of CJ versus ARS in terms of validity, reliability, practicality and acceptability. We recruited two groups of judges/raters, and asked them to use both CJ and ARS to assess English-Chinese consecutive interpreting. In general, we find that ARS had higher concurrent validity, produced more replicable results across rater groups, required less scoring time, and induced higher levels of rater confidence. Additionally, we observe that using either CJ or ARS, the experienced raters tended to generate more valid measures than their novice counterparts and that the assessments of English-to-Chinese interpreting seemed to be more valid than the opposite direction.

The evidence we have obtained so far leans in favor of ARS. Several caveats, however, should be taken into account. First, the assumption that judges should make independent comparisons may have been violated. Given that there were only 20 recordings for each direction, it is possible that judges may have encountered the same recordings in multiple comparisons, thus developing some degree of familiarity with certain recordings (see CJ-D5). Second, although we employed two judge/rater groups of varying scoring expertise, we were not able to recruit expert judges/raters who may produce different assessment outcomes. Third, we did not encourage the judges to use time-saving strategies (for example spot-checking) to expedite CJ. Had such strategies been encouraged, CJ might have been less time-consuming.

In conclusion, the evidence derived from our exploratory study indicates the greater utility of ARS in interpreting assessment, although CJ represents a fairly valid and reliable method. Going forward, we may explore specific assessment conditions in which the effectiveness of ARS, CJ and other scoring methods can be maximized.

ACKNOWLEDGEMENTS

This work was supported by National Social Science Foundation (grant number: 18AYY004).

NOTES

1. https://nomoremarking.com
2. https://www.surveymonkey.com/r/YTNXSQC
3. https://www.surveymonkey.com/r/JM8N3G2
4. Linacre, John M. (2021). A user's guide to FACETS: Rasch-model computer programs. Consulted on June 16, 2021, <https://winsteps.com/a/Facets-Manual.pdf>.
5. In the CJ condition, each rater provided two confidence ratings, one for each direction, whereas in the ARS condition each rater generated a total of six ratings, with three ratings for each direction (that is confidence on InfoCom, FluDel, and TLQual). To make the CJ and ARS conditions comparable, we used a confidence rating, averaged across the three quality dimensions in the ARS condition, to represent raters' confidence level for each direction.

REFERENCES

AIIC (1982): Practical Guide for Professional Conference Interpreters. *AIIC*. Consulted on October 10, 2018, <https://aiic.org/document/547/AIICWebzine_Apr2004_2_Practical_guide_for_professional_conference_interpreters_EN.pdf>.

Andrich, David (1978): Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*. 2:451-462.

Bradley, Ralph A. and Terry, Milton E. (1952): Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*. 39:324-345.

Bramley, Tom, Bell, John and Pollitt, Alastair (1998): Assessing changes in standards over time using Thurstone paired comparisons. *Education Research and Perspectives*. 25:1-24.

Gile, Daniel (1995): Fidelity assessment in consecutive interpretation: An experiment. *Target*. 7(1):151-164.

Grbić, Nadja (2008): Constructing interpreting quality. *Interpreting*. 10(2):232-257.

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

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

Han, Chao (2017): Using analytic rating scales to assess English/Chinese bidirectional interpretation: A longitudinal Rasch analysis of scale utility and rater behavior. *Linguistica Antverpiensia New Series—Themes in Translation Studies*. 16:196-215.

Han, Chao (2018a): Latent trait modelling of rater accuracy in formative peer assessment of English-Chinese consecutive interpreting. *Assessment & Evaluation in Higher Education*. 43(6):979-994.

Han, Chao (2018b): Using rating scales to assess interpretation: Practices, problems and prospects. *Interpreting*. 20(1):59-95.

Han, Chao (2019): A generalizability theory study of optimal measurement design for a summative assessment of English/Chinese consecutive interpreting. *Language Testing*. 36(3):419-438.

Han, Chao, Chen, Sijia and Fan, Qin (2019): Rater-mediated assessment of translation and interpretation: Comparative judgement versus analytic rubric scoring. *The 5th International Conference on Language Testing and Assessment*, Guangdong University of Foreign Studies, Guangzhou, June 6-7, 2019. (Unpublished)

Han, Chao, Chen, Sijia, Fu, Rongbo, *et al.* (2020): Modeling the relationship between utterance fluency and raters' perceived fluency of consecutive interpreting. *Interperting*. 22(2):211-237.

Hartley, Anthony, Mason, Ian, Peng, Grace, *et al.* (2003): Peer- and self-assessment in conference interpreter training. *Centre for Languages, Linguistics and Area Studies (LLAS), University of Southampton*. Consulted on June 10, 2015, <https://researchportal.hw.ac.uk/en/publications/peer-and-self-assessment-in-conference-interpreting-training>.

Jones, Ian and Inglis, Matthew (2015): The problem of assessing problem solving: can comparative judgment help? *Educational Studies in Mathematics*. 89(3):337-355.

Jones, Ian and Wheadon, Chris (2015): Peer assessment using comparative and absolute judgment. *Studies in Educational Evaluation*. 47:93-101.

Jones, Ian, Swan, Malcolm and Pollitt, Alastair (2015): Assessing mathematical problem solving using comparative judgment. *International Journal of Science and Mathematics Education*. 13:151-177.

Laming, Donald (2004): Marking university examinations: some lessons from psychophysics. *Psychology Learning and Teaching*. 3:89-96.

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

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

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

Liu, Minhua (2013): Design and analysis of Taiwan's interpretation certification examination. *In*: Dinna Tsagari and Roelof Van Deemter, eds. *Assessment issues in language translation and interpreting*. Frankfurt: Peter Lang, 163-178.

Liu, Minhua, Chang, Chia-chien and Wu, Shao-chuan (2008): Interpretation evaluation practices: Comparison of eleven schools in Taiwan, China, Britain, and the USA. *Compilation and Translation Review*. 1(1):1-42.

Luce, R. Duncan (1959): Individual choice behavior: A theoretical analysis. New York: Wiley.

McMahon, Suzanne and Jones, Ian (2015): A comparative judgment approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice*. 22(3):368-389.

Mead, Peter (2005): Methodological issues in the study of interpreters' fluency. *The Interpreters' Newsletter*. 13:39-63.

Pollitt, Alastair (2012): Comparative judgment for assessment. *International Journal of Technology and Design Education*. 22(2):157-170.

Pollitt, Alastair and Murray, Neil L. (1996): What raters really pay attention to? *In*: Michael Milanovic and Nick Saville, eds. *Studies in language testing 3: Performance testing, cognition and assessment*. Cambridge: Cambridge University Press, 74-91.

Setton, Robin and Dawrant, Andrew (2016): *Conference Interpreting: A Trainer's Guide*. Amsterdam: John Benjamins.

Steedle, Jeffrey T. and Ferrara, Steve (2016): Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*. 29(3):211-223.

Tarricone, Pina and Newhouse, C. Paul (2016): Using comparative judgment and online technologies in the assessment and measurement of creative performance and capacity. *International Journal of Educational Technology in Higher Education*. 13:1-11.

Thurstone, Louis Leon (1927): A law of comparative judgment. *Psychological Review*. 34:273-286.

Thurstone, Louis Leon (1954): The measurement of values. *Psychological Review*. 61(1):47-58.

Wang, Ji-hong, Napier, Jemina, Goswell, Della, *et al.* (2015): The design and application of rubrics to assess signed language interpreting performance. *The Interpreter and Translator Trainer*. 9(1):83-103.

Wu, Jessica, Liu, Min-hua and Liao, Cecilia (2013): Analytic scoring in interpretation test: Construct validity and the halo effect. *In*: Hsien-hao Liao, Tien-en Kao and Yaofu Lin, eds. *The Making of a Translator: Multiple Perspectives*. Taipei: Bookman, 277-292.

Wu, Shao-chuan (2010): Assessing simultaneous interpreting: A study on test reliability and examiners' assessment behavior. Doctoral thesis, unpublished. Newcastle: Newcastle University.

APPENDIX

Descriptor-based rating scales for assessing consecutive interpreting

Band/Scoring Criteria	**Information Completeness** (InfoCom)	**Fluency of Delivery** (FluDel)	**Target Language Quality** (TLQual)
Band 4 (Score range: 7-8)	A substantial number of original messages delivered (that is, > 80%), with few 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 3 (Score range: 5-6)	Majority of original messages delivered (that is, 60-70%), with only 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 only a few instances of unnatural expressions and grammatical errors.
Band 2 (Score range: 3-4)	About half of the original messages delivered (that is, 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 1 (Score range: 1-2)	A small portion of the original messages delivered (that is, < 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.