Research Methods in Applied Linguistics 2 (2023) 100048

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Halo effects in rating data: Assessing speech fluency

Stefan O'Grady

University of St Andrews, International Education Institute, St Andrews, Scotland

A B S T R A C T

Fluency is a common objective in English language learning and teaching. However, researchers have commented on the absence of a widely accepted definition of the construct and this sense of uncertainty may hinder efforts to measure fluency for purposes of research or assessment. To date, the extent to which rating instruments measure fluency independently from other areas of speech production such as complexity and accuracy has been under-explored. This is a significant gap because the literature broadly suggests that rater scores are susceptible to halo effects that have a distorting influence on the measurement of speaking skills and blur boundaries between assessment criteria. To investigate this issue, the current study examines a data set of scores assigned to 77 English language learners on two speaking tasks using an analytic rating scale featuring criteria for speech complexity, accuracy and fluency (CAF). The tasks were transcribed and analysed using measures of CAF. Rater scores were analysed using many-facet Rasch measurement and multiple regression. Results revealed that rated fluency was influenced by lexical complexity, indicating that fluency scores represented more than the fluency construct outlined in the analytic scale. Measures of speech speed, phonation time ration, length of utterance, lexical complexity, total speaking time and repair fluency explained the largest amount of variance in the fluency scores. Implications for research, language teaching and assessment are discussed.

1. Introduction

Attainment of second language fluency has the potential to determine future educational, economic and professional opportunities (Chou, 2018; McCarthy, 2010). For this reason, fluency judgements are a central component in language teaching and assessment. However, the process of assigning scores to speech is complicated by an absence of approved standards or a commonly accepted definition of fluency (De Jong, 2018; Tavakoli & Hunter, 2018). In applied linguistics, spoken fluency is commonly investigated as a component of the complexity, accuracy and fluency measurement framework (CAF; Ellis et al., 2019; Housen et al., 2012). CAF is an analytical approach to speech measurement that examines audio recordings and transcripts to determine levels of grammatical and lexical accuracy and complexity in addition to fluency, which is expressed in terms of speech speed, breakdown, and repair (Skehan, 2009). CAF facilitates comparisons between individuals or groups and has commonly been applied to contrast pedagogical interventions such as extending the amount of planning time language learners are permitted before a speaking task, which has consistently demonstrated effects on fluency but only marginally on complexity and accuracy (Qin & Zhang, 2022). Ellis et al. (2019) have recently argued that the CAF framework may inform the development of rating scales for teachers, researchers, and examiners to measure spoken fluency. Crucially, this interpretation assumes that raters working with a rating scale will distinguish between analytic criteria in a way that resembles a researcher working with a transcription. This assumption is problematic because rater effects, most notably halo effects that distort the intended distinctions between separate analytic criteria, are common in speaking score data (Yorozuya & Oller, 1980). Essentially, the perspective that fluency is independent from complexity and accuracy may simply not correspond to the way raters assess the construct. The motivation for the current study is therefore to investigate the argument made in Ellis et al. (2019) by examining the application of an analytic rating scale featuring descriptors for fluency in addition to complexity and accuracy.

E-mail address: so59@st-andrews.ac.uk

https://doi.org/10.1016/j.rmal.2023.100048
Received 2 December 2022; Received in revised form 8 March 2023; Accepted 8 March 2023
2772-7661/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

2. Literature review

Although spoken fluency frequently appears as an objective in language teaching syllabi, as a pedagogical construct, the meaning of fluency is difficult to pin down. Popular perceptions are interchangeable with language proficiency or competence (McCarthy, 2010). However, in applied linguistics fluency represents a narrower, phonological construct that reflects the levels of automaticity of the cognitive processing involved in speech production (Levelt, 1989; Pawley & Syder, 1983; Segalowitz, 2010; Tavakoli & Hunter, 2018). In a widely cited introduction to second language spoken performance, Housen et al. (2012), p.2) outline a theoretical framework in which fluency is defined as 'the ability to produce the L2 with native-like rapidity, pausing, hesitation, or reformulation' and is principally distinguishable from the accuracy and complexity of language use. Within this framework, fluency is evaluated based on phonological analysis of speech speed, breakdown and repair (Skehan, 2009; Tavakoli & Wright, 2020). Phonological analysis has been shown to account for variation in speaking test scores (Kahng, 2018), and features in automated scoring systems of second language speech (Zechner & Evanini, 2020).

In the research literature, the common consensus is that this kind of phonological analysis represents "empirical ratification of how fluency is realised in real language use" (McCarthy, 2010, p.13). Fluency measures generated from phonological analysis have been shown to be heavily impacted by characteristics of the speaking task and task conditions such as the level of familiarity speakers have with the topic and the opportunity to plan their speech in advance (Bui & Huang, 2018). On the basis of this research, recommendations may be made about pedagogical and assessment task design in a way that fluency may be systematically raised as a product of specific task characteristics (Ellis et al., 2019). However, teachers and examiners are unlikely to judge task fluency with phonological analysis and it is widely recognised that the process of raters assigning scores to speech is largely idiosyncratic (Lumley, 2002; Pill & Smart, 2020). Therefore, to investigate the Ellis et al. (2019) argument that CAF may also be used in rating scale design, validation research is crucial to determine whether raters view fluency in the same way that the CAF measures conceive of the construct. Fulcher (2015, p.76) suggests this is unlikely because fluency in CAF does not reflect the influential effect of contextual factors and the surrounding discourse: 'what then can be the purpose of simply counting pauses, or measuring pause length or speech rate, when these vary for a variety of reasons, only some of which are related to L2 language proficiency?' This interpretation implies that raters may be inclined to assess fluency more holistically as part of a contextually embedded judgement that considers aspects such as the complexity and accuracy of the speech. Ultimately, this argument undermines the assumption Ellis et al. (2019) make that rating scales can be used in a similar way as CAF measures.

Raters of speaking assessments are typically required to perform multiple operations simultaneously, often acting as interlocutor while processing the contents of a rating scale to award a score to speech as it unfolds. In cases where the rater is removed from the spoken event, such as when the task is scored from a recording, allocating similar amounts of attention to the separate components of an analytic rating scale may be unnatural and impose a substantial cognitive burden on the rater with the result that seemingly separate scores may reflect a mishmash of different traits and constructs (Knoch & Chapelle, 2018). Essentially, the cognitive load associated with analytic scale use may undermine the potential benefits to provide fine-grained descriptions of performance (Baddeley, 2007; Green, 2021). When raters consistently assign similar scores to separate traits on an analytical scale it may be due to a combination of natural correlations between criteria describing different dimensions of the same construct and difficulty discerning between the criteria of the scale. In each situation, the ratings are influenced by a halo effect (Myford & Wolfe, 2003, 2004). The former is a consequence of dividing a construct or latent trait into component parts in a rating instrument and is referred to as a true halo effect (Eckes & Jin, 2022), whereas the latter is representative of "cognitive distortions, errors in observation and judgment, and rating tendencies of the individual rater" (Murphy et al., 1993, p. 220). In essence, halo effects in fluency scores may be directly attributable to the raters' appraisal of accuracy or complexity. Halo effects prevent reliable interpretation of scale scores and obscure decisions taken or feedback provided on their basis (Eckes, 2011). Fisicaro and Lance (1990) suggest that halo effects may be caused by a first impression of the examinee that influences subsequent scoring of different criteria (e.g. age, gender, ethnicity), misunderstanding the intended differences between criteria (e.g. owing to a lack of training), or excessive emphasis on a single feature of the task performance (e.g. accent, pitch, volume). For instance, if a rater is preoccupied with the grammatical accuracy a student displays in the completion of a speaking task, this preoccupation may carry over into different scale criteria in a way that the accuracy score determines, for example, the fluency score. In recent research, Kim (2020) observes that raters are most likely to demonstrate a halo effect when scale criteria that are conceptually similar are presented successively (i.e. organisation presented closely to content, and vocabulary presented next to language use).

In one of the few studies to investigate halo effects in a speaking task, Yorozuya and Oller (1980) report the results of correlation and factor analysis demonstrating halo effects between analytic scale criteria of vocabulary, grammar, fluency, and pronunciation. The researchers conclude that halo effects play an influential role in speaking assessment and analytic scales are more likely to represent the raters' evaluation of overall communicative effectiveness than the separate criteria in the scale. More recently, researchers have applied many facet Rasch measurement (MFRM) to detect halo effects in score data (Farrokhi & Esfandiari, 2011; Kim, 2020; Knoch et al., 2007; Kozaki, 2004; Schaefer, 2008). MFRM is particularly well suited to examine halo effects because the statistical approach is designed to identify unexpected rating patterns that do not fit the predictions of the Rasch model (Eckes, 2011). MFRM may detect sizeable overall halo effects whereby all raters display a clear halo, or smaller effects at a level whereby the effect is observable in an individual's rating or among a small group of individuals (Kim, 2020). Research studies investigating halo effects with MFRM predominantly report halo effects at the individual level and this indicates that detecting halo effects may require scrutinisation of individual rater scores (Farrokhi & Esfandiari, 2011; Kim, 2020; Knoch et al., 2007). However, this research examined tests of second language writing ability and to date, research has not explored halo effects in speaking assessments using MFRM. Furthermore, the extent to which complexity, accuracy and fluency measures may account for halo effects in test data is currently uncertain. This is an

important focus because CAF measures have potential to identify the characteristics of speech raters prioritise when assigning scores and hence provide further insight into the ways raters conceive of fluency.

3. Research questions

The research to date has not investigated halo effects in speaking tests using many-facet Rasch measurement or the potential to account for variation in fluency scores with measures of CAF. This represents a substantial gap in the literature as halo effects between different criteria are a common feature of rating scale scores (Myford & Wolfe, 2004) and the literature suggests that raters may find it difficult to separate the fluency of an utterance from other aspects of the discourse (Fulcher, 2015). Halo effects may indicate that fluency scores have been influenced by accuracy and complexity, and may be large enough to affect the entire test or isolated and affect scores from a single rater or small group of raters. To address these gaps, the current study reanalysed a dataset featuring scores on a rating scale of complexity, accuracy, and fluency to answer the following research questions.

1 Is there a discernible halo effect overall between rating scale criteria for complexity, accuracy, and fluency in the scores of a second language speaking assessment task?
2 If an overall halo effect is not detected, does analysis at the individual level identify raters displaying a halo effect?
3 If a halo effect is uncovered overall or at the individual level, what combination of CAF measures best accounts for the variation in fluency scores?

4. Method

4.1. Participants

Students. The participants were 77 Turkish learners of English that had been studying in the English language preparatory programme of an English-medium university for one year. At the time of the study, participants were enrolled in a summer school program designed to prepare them for the English language university entrance exam, which assesses the candidates' ability to follow English-medium instruction at the undergraduate level. This is estimated by the university administration as the 'B2' level on the Common European Framework of Reference for Languages (Council of Europe, 2001). The participants' age range was between 18 and 25. All participants signed letters of consent informing that they were taking part in research intended to assess the impact of changes to institutional assessment formats.

Raters. Fifteen instructors of English representing the range of language backgrounds and teaching experience in the institution took part in the study. Seven of the instructors were native speakers of English and the remaining eight were native speakers of Turkish. The range of their teaching experience was from five to 25 years (mean = 12.39, SD = 5.76). The instructors regularly worked as examiners in institutional speaking exams and were responsible for completing formative assessments of speaking ability in their own English classes.

4.2. Procedures

4.2.1. Tasks

Two pedagogical tasks were used to elicit speaking samples in the study. The tasks had previously been used in the institution as part of the speaking section of the university entrance English language test but had been retired and were being used as formative assessment material at the time of this study. The students had not completed these specific tasks previously, although they were familiar with the format. The study took place in the participants' language classrooms, with the researcher acting as the interlocutor. Before the tasks began, participants first completed a warmup session during which they answered a series of personal questions. The warmup session was not included in the analysis. After the warmup session had been completed, participants completed two monologue tasks with the following instructions:

a) Describe something interesting you have recently heard in the news.
b) Describe an experience that changed your life.

The task performance was recorded using Audacity (2.0.6, 29 September 2014, http://audacity.sourceforge.net) and saved both as Audacity files and in MP3 format. Audacity files include a visual representation of the waveform that facilitates transcription and transcription was completed from the Audacity files. The MP3 files were distributed to the raters for scoring.

4.2.2. Analytic rating scale

An analytic rating scale developed by Iwashita et al. (2001) was selected for the rater analysis. The scale contains descriptors for five band levels of proficiency in complexity, accuracy and fluency that range from beginner to advanced language user and is particularly suitable for examining associations between measures of CAF and rating scale scores (see Appendix 1). The scale has featured in several research studies and has the advantage of being previously validated for research (Elder & Iwashita, 2005; Elder et al., 2002; Nitta & Nakatsuhara, 2014).

4.2.3. CAF measures

Various operationalizations of speech fluency have been investigated in the literature (De Jong, 2018; Pallotti, 2020; Suzuki & Kormos, 2020). Of these, many measures are calculated with a simple acoustic analysis of a sound file and a transcription. For example, speech rate is taken to indicate the speed with which speech is produced and is evaluated by establishing the total number of syllables or words in a transcription and dividing this by the total speaking time (Ginther et al., 2010). Phonation time ratio measures breakdown fluency by comparing the total amount of time spent pausing and hesitating with the total amount of time spent speaking (Suzuki, 2021). Measures of breakdown in spoken fluency provide insights into the underlying processing challenges speakers experience when formulating and producing second language speech (Tavakoli et al., 2020). Breakdown in fluency may occur both between (i.e. pauses) and within (i.e. hesitations) syntactic boundaries (Field, 2011). Whereas pauses allow speakers to generate content (i.e. pauses facilitate speech conceptualization), hesitations are indicative of increased effort involved in the completion of an utterance (i.e. hesitation facilitates speech formulation; Skehan et al., 2016). Excessive hesitation is assumed to be due to gaps in language proficiency and is more likely to impact an interlocutor's impression of fluency than excessive pausing (Préfontaine, 2013). Pauses and hesitations may involve periods of silence or may be filled with fillers such as *erm, um,* or *mmm*. Filled pauses and hesitations are very common in spoken discourse and serve a similar purpose as unfilled pauses and hesitations (Kormos, 2006). However, the literature indicates that filled and unfilled pauses and hesitations may be perceived differently; the filled form is often interpreted as a speaker's intention to continue a conversational turn (Levelt, 1989). Distinguishing between pauses and hesitation permits further detailed phonological analysis of the speech such as mean length of utterance, a measure of the number of words produced between filled/unfilled pauses or hesitations. The ability to produce speech without having to pause or hesitate is a key indicator of the extent to which the processes of language retrieval and encoding of the speech have become proceduralised and is symptomatic of the speakers' current state of language proficiency in the second language (Field, 2011). The more words a speaker can produce without having to hesitate, the more fluent they appear. An additional measure that has been explored in the literature is repair fluency, which involves a calculation of the proportion of errors in language or content that the speaker reformulates and measures the extent to which the speaker is consciously monitoring their speech for accuracy (Pallotti, 2020).

Transcription of the speech followed Fulcher and Davidson (2007); Jefferson (2004) and examples presented in Foster et al. (2000). To run syntactic analysis, transcripts were divided into AS-units: a unit developed specifically for the analysis of speech that involves a 'single speaker's utterance consisting of an independent clause or sub-clausal unit, together with any subordinate clause(s) associated with either' (Foster et al., 2000, p.365). CAF measures identified in De Jong (2018) were evaluated to establish how researchers had measured complexity, accuracy and fluency in the literature. As the focus of the current study was on fluency scores, a large number of fluency measures were required to measure the various aspects of the construct (Tavakoli et al., 2020). In contrast, it was deemed sufficient to utilize general measures of accuracy and complexity to determine whether these constructs had any impact on the fluency scores. Complexity is conventionally divided into measures of syntactic and lexical complexity (Skehan, 2009). The following CAF measures were identified (studies applying these measures are indicated in parenthesis):

4.2.3.1. *Fluency*.

The frequency of hesitations was assessed through the mean number of hesitations per AS-unit. The calculation involved establishing the number of hesitations and dividing by the number of AS-units (Skehan & Foster, 2005).

The proportion of task time the speaker spent producing speech was assessed through phonation time ratio. The calculation involves establishing the amount of time spent producing speech and subtracting the total amount of time spent pausing. This figure is divided by the total amount of time spent completing the task and expressed as a percentage (Bui & Huang, 2018).

The proportion of pauses and hesitations that were filled was assessed by calculating the number of filled hesitations/pauses and dividing by the total number of hesitations/pauses; the resulting figure was multiplied by 100 and expressed as a percentage (Skehan & Foster, 2005).

The ability to produce speech without having to pause and hesitate was assessed through mean length of utterance. The calculation was the total number of words divided by the total number of pauses and hesitations. This was expressed as the average number of words between pauses and hesitations (Li et al., 2014).

The amount of time used to complete the task was expressed in seconds as the total speaking time.

Speech speed was assessed though speech rate. The calculation involved the number of words divided by speaking time, multiplied by 60 and expressed as words per minute (Li et al., 2014).

Repair fluency was measured by calculating the percentage of errors in syntax, lexis or content that were reformulated (Nitta & Nakatsuhara, 2014).

4.2.3.2. *Complexity*.

Lexical diversity was assessed using Guiraud's Index. This was calculated by establishing the type-token ratio of each transcript and dividing the type value by the square root of the token value (Gilabert, 2007).

Syntactic complexity was assessed through the mean number of clauses per AS-unit. This was calculated by counting the number of dependent and independent clauses in each transcript and dividing by the number of AS-units (Skehan & Foster, 2005).

4.2.3.3. *Accuracy*.

Frequency of grammatical and lexical errors was assessed through the mean number of errors per AS unit. This was calculated by counting the number of errors in each transcript and dividing by the number of AS-units (Li et al., 2014).

Pauses and hesitations were defined as a period of silence in excess of one second (Foster & Skehan, 1996). Although a common criterion in SLA research for pauses and hesitations is .25 seconds (Kormos, 2006; Tavakoli & Wright, 2020), instances of pauses of this length were very frequent in the sound files, often occurring between every word, and during training the raters agreed that such instances could not be regarded as genuine pauses. It was imperative for the CAF measures to reflect as closely possible the raters' interpretations of fluency as the research aim was describing variation in their scores. However, the adoption of the one second standard diverges from the criterion established in the literature (e.g. Tavakoli & Wright, 2020) and may hinder comparisons between the current study and those that have come before. All instances of pauses and hesitations were identified through analysis of the waveform provided in the *Audacity* program (2.0.6, 29 September 2014, http://audacity.sourceforge.net). Reliability was evaluated by having a second transcriber transcribe a proportion of the total: 10 per cent of the data was transcribed and coded for all measures. Inter-coder reliability (TOTAL AGREEMENT/ n x 100) was 86.25 per cent.

4.2.4. Rating process: analytic scale

Before scoring the samples, the raters took part in a standardisation session. Prior to the session, three task recordings were identified by the researcher as representative of the range of abilities in the sample of students. During standardisation, the raters first analysed and discussed the contents of the rating scale with the researcher to ensure that a common interpretation of the content was reached. Following this, the student samples were presented to the group. The raters discussed the fluency, accuracy and complexity of the samples with reference to the scale content and the researcher highlighted aspects of the samples that exemplified the content of specific complexity, accuracy, and fluency levels. Once appropriate scores had been established for each standardisation sample, raters then independently scored between 20 and 99 samples (mean = 29.1, SD = 19.1). The raters were matched multiple times by assigning scores to the same students to ensure sufficient connectivity in the data to run statistical analyses (see Appendix 2. Rater Matrix).

4.2.5. Statistical analytical procedures

The scale ratings were analysed using the Facets program, software for running many facet Rasch measurement (3.71.4, 18 January 2014, www.winsteps.com) that has been identified as suitable for detecting halo effects (Eckes & Jin, 2022; Kim, 2020). Facets calibrates the variables under investigation to the same interval scale of the latent trait, the logit scale, so that all variables share the same measurement. Three facets were entered into the analysis; test taker, rater, and scale criteria (complexity, accuracy, and fluency), using a rating scale model that assumes raters interpret and apply the rating scale in the same way (Eckes, 2011). Scores were assigned to both tasks but task was not entered as a facet in the model because the focus was on overall halo effects rather than halo effects on specific tasks. The equation for the model was

$$log(P_{njik} / P_{njik-1}) = B_n - R_j - D_i - F_k$$

where P_{njik} = probability of receiving rating k under the following circumstances, $P_{njik\text{-}1}$ = probability of k-1, B_n = ability of student n, R_j = severity of rater j, D_i= difficulty of scale criteria i, F_k = difficulty of category k relative to k -1

The analysis generates fit statistics for all facets; infit and outfit statistics range from 0 to infinity and indicate the extent to which the data conform to the expectations of the Rasch model. In the literature, commonly proposed fit acceptability values range from .70 to 1.30; values in excess of 1.30 indicate underfit and that the data is unpredictable, e.g. a rater does not demonstrate a clear and consistent pattern of severity when scoring, whereas values below .70 indicate overfit and that the data is overly consistent, e.g. a rater is conservative and does not use the full range of the scale (Bonk & Ockey, 2003; Myford & Wolfe, 2004). Critical values of .70 to 1.30 are practical guidelines, but fit statistics are continuous variables, and some researchers suggest using empirical methods to set sample specific critical values (Guo & Wind, 2021). Overall, values that fall within a range of 0 to .70 or 1.3 to 2.00 are not considered constructive for measurement but also do not have a distorting effect and may thus be retained in the analysis (Linacre, n.d.). Outfit is sensitive to outliers that may inflate values, whereas infit is indicative of patterns in the data (Linacre, 2019). Facets also assigns measure values in logits to each variable for comparison and generates fair average scores for each test taker, rater and scale criteria. Assuming acceptable fit to the model, the fair average represents a contextualised average score on each criterion that is unaffected by the idiosyncrasies of each rater or any disproportionate allocation of advanced/weak students to each rater. In this study, the fair average represents a composite of scores assigned to the two tasks. Initial analysis demonstrated that four of the students were associated with infit statistics that exceeded 2.00. In such cases, Linacre (n.d.) recommends removing the misfitting data and running the analysis to create an anchor file to preserve the data that fit the model. The anchor file and misfitting data were combined and on the basis of this measurement, the reliability of the data and the fit of the data to the Rasch measurement model was established using reliability of separation statistics and mean square fit statistics. In addition, comparisons between the measure values of the analytic rating scale criteria were made. Following Myford and Wolfe (2004), an alternative method for detecting halo effects at the individual rater level was followed. The approach involved anchoring the separate criteria of the analytic rating scale to the same level of difficulty and running an analysis to determine whether any of the raters fit the adjusted model using fit mean square statistics. Those raters found to fit the adjusted model are likely to be exhibiting a halo effect because the lack of variability in their scores matches the lack of variability in the adjusted model: "The idea is to explicitly match the model used in the analysis to the aberrant rater behavior the researcher wants to detect" (2004, p. 209). The results of this analysis were evaluated to identify raters that fit the adjusted measurement model and therefore exhibited a halo effect.

To answer research question three, it was necessary to generate fluency fair average values for each test taker that were not impacted by the complexity and accuracy data, for example by affecting fit statistics. A second model was created using the same equation as above but with fluency scores entered as the only criterion (Eckes, 2011). Misfitting students associated with infit values

Table 1
Scale criteria measure values.

Criteria	Fair Average	Measure	SE	Infit MS	Outfit MS	Fixed (all same) chi-square
Fluency	2.71	-.48	.08	1.24	1.26	χ^2=
Accuracy	2.53	.00	.08	1.06	1.08	66.9,
Complexity	2.36	.49	.08	1.08	1.03	p = .00

above 2.00 were again identified in this analysis (n=6) and the same anchoring procedure was followed to preserve the fitting values. The results of the second anchored analysis were evaluated for a halo effect using stepwise multiple regression analysis to evaluate the relative contributions of the various CAF measures on the fluency fair average scores. This involved entering fair average fluency values as the dependent variable and the fluency measures as the predictor variables into the regression model, and in a stepwise approach adding the complexity and accuracy measures to determine whether the explained variance would significantly increase (Plonsky & Ghanbar, 2018). The sample size of 77 was deemed sufficient to run multiple regression involving the CAF measure predictor variables (VanVoorhis & Morgan, 2007). De Jong (2018) raises concerns specifically about the multicollinearity of different fluency measures. Multicollinearity was investigated using VIF and tolerance values, which did not indicate the presence of multicollinearity (Plonsky & Ghanbar, 2018). Initial analysis indicated that the data met the assumptions for conducting multiple regression.

5. Results

The results of the Facets analysis are first presented in the form of a Wright map (see Fig. 1). The Wright map orders the individual elements of each facet on the logit scale (the first column) by placing higher scoring test takers (the second column) and more severe raters (the third column) toward the top of the map. The rating scale criteria (the fourth column) are organised with more difficult criteria appearing toward the top of the map. The final column on the map represents the rating scale. In this column, band level five is presented in parenthesis, indicating that it was rare for test takers to be awarded this score. The map demonstrates that test taker ability estimates were spread between -5 and 3 logits on the logit scale, indicating that there was a wide range of speaking ability in the population. The spread of test taker ability exceeds the range of rater severity; however the fifteen raters clearly demonstrate different levels of severity when assigning scores. The raters were separated into 6.45 statistically distinct levels of severity and the reliability of the separation was .98 (Linacre, 2019).

To answer research question one, the rating scale criteria are first examined; raters were most severe when assigning complexity scores and most lenient when assigning fluency scores. This finding is consistent with research that has used the same scale, which also found that fluency was most leniently scored (Nitta & Nakatsuhara, 2014). The distinct location of criteria on the map indicates that the raters were generally able to discern between the criteria when assigning scores and interpreted the scale criteria as separate constructs, thereby offering evidence against the presence of a halo effect (Myford & Wolfe, 2003, 2004). Correlations between the subscales were as follows: complexity and accuracy (r = .77, p = .01), fluency and accuracy (r = .69, p = .01), and fluency and complexity (r = .73, p = .01). The independence of the criteria on the analytic scale is further verified by the detailed measure statistics presented in Table 1, which presents the fair average, measure values, standard errors and fit statistics associated with criteria on the analytic rating scale.

The rating scale criteria fair average values ranged from 2.36 to 2.71, which corresponds to a range of .49 to -.48 on the logit scale. The criteria were separated into 4.62 statistically distinct levels (strata 6.49) and the reliability of separation (a statistic corresponding to Cronbach's alpha; Linacre, 2019) was high at .96, indicating that raters distinguished successfully between the criteria. The result of the fixed chi-square test indicates that the difference between the measure values was significant at $p < .001$. Collectively, these results imply that as a group, the raters were not influenced by a halo effect when assigning scores to the speech samples (Myford & Wolfe, 2004).

In answer to research question two, the results of the individual rater analysis are presented in Table 2. The table presents the fair average, measure values, standard errors and infit and outfit mean square statistics associated with each rater using the adjusted measurement model with all scale criteria anchored at 0 logits. The table also includes the results of the fixed (all same) chi-square test, which reports that the difference between rater measure values was statistically significant at $p < .001$. Rater data associated with infit and outfit values falling within a range of .70 to 1.30 exhibit fit to the measurement model expectations. Applying these criteria, raters 1, 2, 3, 5, 7, 8, 12, 13, 14, 15 fit the adjusted model (outfit is particularly sensitive to outliers and the infit statistic associated with rater 5 warrants investigation; Bonk & Ockey, 2003; Myford & Wolfe, 2004). This is strong evidence of a halo effect in the scores these raters provided.

To further investigate halo effects in individual raters' scores, Myford and Wolfe (2004) recommend examining the scores that raters assigned for patterns in the data. Table 3 presents the percentages of sequences of scores provided by the raters fitting the adjusted measurement model that contain identical values between scale criteria. The table shows that the vast majority of score sequences contain evidence of some form of a halo effect. For raters 1, 2, and 7 it was more common for a halo effect to appear between all three criteria, whereas raters 3, 5, 8,12,13,14, and 15 were more likely to exhibit halo effects between two of the criteria. The two most commonly occurring identical sequences were accuracy and complexity, which accounted for 51% of the identical scores between two criteria, followed by fluency and accuracy (26%) and fluency and complexity (23%).

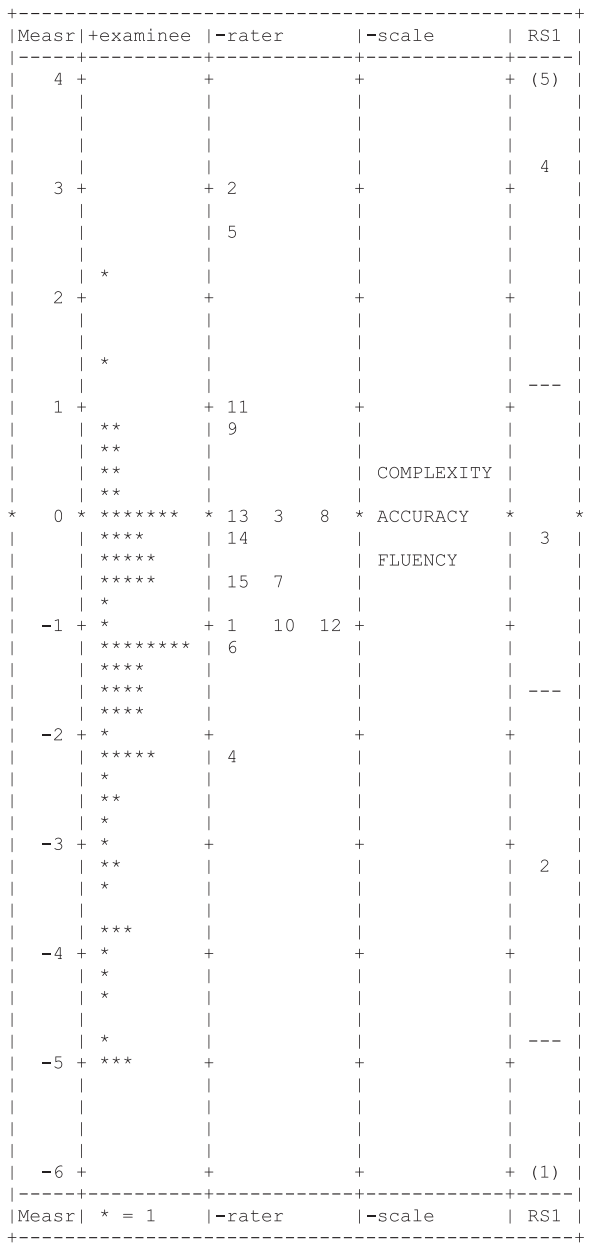

Fig. 1. Wright map.

To answer research question three, multiple regression analysis was completed to evaluate the extent to which the CAF measures, the predicting variables, predicted the fair average fluency scores, the dependent variable (see Table 4). The results of the regression show that six variables accounted for approximately 54% of the variation in fair average scores. These were the mean length of utterance, speech rate, phonation time ratio, total speaking time, Guiraud's index, and repair fluency. The remaining CAF measures did not make a statistically significant contribution to the variation in the fair average measures.

6. Discussion

The current study was conducted to investigate the relative independence of second language fluency ratings. Fluency is conventionally assessed by applying rating scales and the literature indicates that scores may be sensitive to rater effects, such as halo effects (Myford & Wolfe, 2003, 2004), which obscure the measurement of the construct (Yorozuya & Oller, 1980). The results of the MFRM of rater scores demonstrated that overall, raters were capable of separating fluency from complexity and accuracy. Halo effect analyses involving the entire group of raters indicated that there was no overlap between the criteria of the analytic scale.

Table 2
Rater statistics.

Rater	Fair Average	Measure	SE	Infit MS	Outfit MS	Fixed (all same) chi-square
1	2.91	-1.04	.17	0.76	0.75	χ^2=
2	1.60	3.08	.28	1.15	1.14	481.7,
3	2.51	0.07	.19	0.73	0.74	p = .00
4	3.30	-2.12	.22	1.51	1.58	
5	1.73	2.63	.27	1.15	1.49	
6	2.96	-1.17	.22	2.82	2.79	
7	2.72	-0.51	.10	0.91	0.91	
8	2.50	0.08	.20	1.11	1.11	
9	2.26	0.79	.21	1.35	1.35	
10	2.90	-1.02	.20	1.55	1.52	
11	2.18	1.07	.21	1.87	1.85	
12	2.91	-1.04	.20	1.00	0.99	
13	2.54	-0.04	.19	1.24	1.23	
14	2.62	-0.25	.21	0.89	0.88	
15	2.73	-0.55	.21	0.83	0.81	
Mean	2.56	0.00	.20	1.27	1.28	

Table 3
Breakdown of score sequences.

	Sequences containing identical scores	Two identical scores*	Three identical scores
Rater 1	100%	44%	56%
Rater 2	100%	30%	60%
Rater 3	100%	62%	38%
Rater 5	90%	60%	40%
Rater 7	100%	49%	51%
Rater 8	96%	58%	38%
Rater 12	96%	54%	42%
Rater 13	96%	59%	37%
Rater 14	100%	65%	35%
Rater 15	100%	65%	35%

* 51% accuracy and complexity (n=93), 23% fluency and complexity (n=42), 26% fluency and accuracy (n=47)

Table 4
Results of the regression analysis.

Model	Predictors*	R	R^2	R^2 Change	F Change	df	p	Durbin-Watson
1	MLU	.442	.195	.195	36.893	1.52	.00	
2	MLU + SPR	.571	.326	.131	29.335	1.51	.00	
3	MLU + SPR + PTR	.654	.428	.102	26.640	1.50	.00	
4	MLU + SPR + PTR + TOTAL	.694	.481	.054	15.407	1.49	.00	
5	MLU + SPR + PTR + TOTAL +REP	.708	.502	.020	6.027	1.48	.02	
6	MLU + SPR + PTR +TOTAL +REP+G.IN	.732	.535	.033	10.574	1.47	.00	1.045

* MLU: Mean length of Utterance; SPR: Speech Rate; G.IN: Guiraud's Index; PTR: Phonation Time Ratio; TOTAL; Total Speaking Time: REP; Repair Fluency

The absence of the halo effect at this level contradicts findings in Yorozuya and Oller (1980), who report a clear halo effect in an analytic scale featuring criteria for grammar, vocabulary, fluency and pronunciation. However, this may be due to differences in the statistical analytical approach adopted in the studies. Whereas Yorozuya and Oller (1980) identified the halo effect using factor analysis, the current study applied MFRM and detected halo effects at the individual rater level. The finding that halo effects were apparent at the individual level confirms results in MFRM studies focussing on writing assessments (Farrokhi & Esfandiari, 2011; Knoch et al., 2007). As in these studies, the halo effect was not sufficiently substantial to impact test level measures such as criteria severity, separation and strata, but was clear in individual measures such as the fit statistics and overlap percentages. The findings of the current study suggest that ten of the raters may have had difficulties disentangling the fluency with which students produced speech from the relative accuracy and complexity that their speech contained and this would seem to corroborate the argument that valid measurement of fluency should also account for linguistic content (Fulcher, 2015).

In high stakes assessments, this kind of halo effect in scores would discredit the validity of any decisions based on test results and constitute a sound basis for a legal challenge of test results (Shohamy, 2001). In formative assessment or task-based research, this mixing of criteria on an analytic scale affects the potential for scores to serve a useful function to form the basis for any future pedagogical decisions. It also mitigates the potential to support learning with assessment; misplaced negative feedback can have a

detrimental effect on student motivation that has the potential to determine a students' ultimate success as a language learner (Muñoz & Ramirez, 2015). Assessment is an important component of language learning programmes that helps students self-regulate, increase self-awareness and set learning goals (Leung & Mohan, 2004; Xiao & Yang, 2019). Failure to identify and adequately define the target of assessment (i.e. the construct) has potential to misguide instruction and lead to erroneous decisions that could impact on students' future learning experiences.

Multiple regression analysis was completed to identify the aspects of the students' speech that account for the variation in fluency scores and to provide further explanation of the halo effect. The results of the multiple regression analyses indicated that the mean length of utterance and speech rate accounted for the largest amount of score variance. This finding corresponds to results reported in the literature that composite measures such as mean length of utterance and speech rate are strongly associated with fluency judgments (Bosker et al., 2013; Riggenbach, 1991; Suzuki et al., 2021). However, in addition to the phonation time ratio, total speaking time and repair fluency, Guiraud's Index, a measure of lexical density, also accounted for a proportion of the score variance. In the context of a recent meta-analysis of studies examining the relationship between fluency measures and fluency scores (Suzuki et al., 2021, p. 451), the predictive power of the regression model involving a complexity measure in addition to fluency measures is strong and confirms speculation that unexplained variance in fluency scores using fluency only measures may be due to raters attuning to "linguistic aspects such as lexis, grammar, and pronunciation". This finding suggests that the raters interpreted fluency not just as a temporal construct akin to speech speed and breakdown but also as embedded in lexically rich discourse. Interestingly, no score variance was explained by the errors per AS unit despite the evidence of a halo effect between accuracy and fluency scores. It may be the case that raters were impacted by inaccuracies that were not identified by the accuracy measure, such as mispronunciation, which has been shown to impact the perceived fluency of an individual's speech (Browne & Fulcher, 2017; Suzuki & Kormos, 2020).

Overall, these results indicate that in this study fluency scores represent an ability to produce the L2 with speed and limited breakdowns, correct errors and speak with a variety of lexis. The findings suggest that future research seeking to relate human judgements of fluency to objective measures of fluency may also need to investigate the relative contribution of measures of complexity and accuracy. This finding has implications for measurement of fluency and provides a different perspective on the argument that CAF based rating scales may provide similar data as measures of CAF (Ellis et al., 2019). For practitioners, assessment of spoken fluency using rating scales is quick and simple to conduct. However, an administrative decision may need to be made about whether to interpret fluency as a distinct construct divorced from the linguistic and propositional aspects of speech and make efforts to train raters to prevent future halo effects, or to adopt a more holistic interpretation of fluency. Based on the results of this study, a more holistic interpretation of fluency might incorporate aspects of speech complexity and an appropriate band level descriptor for fluency would include reference to speed, breakdown, and lexical variety. However, this finding may reflect a very contextualised interpretation of the construct and more research in different contexts is required to verify this claim. Fluency scale construction may also draw upon teacher accounts of the way fluency manifests in their students' speech during formative assessment and classroom interaction (Turner & Upshur, 1996). In this case, if the inseparability of fluency from areas such as accuracy and complexity emerges as a genuine concern, the theoretical framework associated with analytic scales may need to be reconsidered for speaking assessment. For researchers working with analytic scales, the results of the current study indicate that it is important to examine data for halo effects and a combination of many facet Rasch measurement of scores and analysis of test samples using an analytical framework such as CAF has potential to both identify and account for these effects.

An important limitation of the study is the specific focus on participants' fluency in monologue tasks. In language learning and teaching, interaction is crucial and fluency is collaborative (An et al., 2021; Lo & Macaro, 2015; Peltonen, 2017). Future research may therefore explore the assessment of fluency in co-constructed discourse between students, and between students and teachers. It may also be necessary to examine halo effects with rating scales that have been developed for specific tasks in specific contexts as in the current study any gaps between the scale contents and fluency measures may have attenuated relationships between performance and scores. Future research may also investigate the relationship between fluency and relevant constructs such as coherence, cohesion, pronunciation and intelligibility, which may impact on fluency judgements in language assessments (Browne & Fulcher, 2017). Finally, it may be important to investigate associations between rater characteristics such as experience, first language background, and proficiency in the target language, and halo effects in speaking tests to inform future rater training programs (Zhang & Elder, 2011). Qualitative data in the form of rater accounts may prove informative in this respect (Isaacs & Thomson, 2020).

7. Conclusion

The present study examined the independence of fluency scores on a second language speaking task by investigating halo effects in rater score data and conducting CAF analysis. The results are most likely to be of interest to teachers and institutions engaged in speaking research and assessments. However, the results are also generalisable to high stakes language assessments involving rating scales and specifically to test developers wishing to encourage a better understanding of score meaning among stakeholders. In tests of second language speaking, fluency scores may reflect more than temporal features of the speech such as speed and frequency of pauses and extend to include aspects of speech content such as accuracy or complexity (Pallotti, 2009, 2020; Tavakoli & Wright, 2020). In this sense, fluency scores based on the contents of rating scales may closer reflect the more popular interpretation of fluency as a proxy for proficiency than the specified construct studied in applied linguistics (Ellis et al., 2019; McCarthy, 2010; Yorozuya & Oller, 1980).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.rmal.2023.100048.

Appendix 1. Analytic rating scale (Iwashita et al., 2001)

Fluency

5. Speaks without hesitation; speech is generally of a speed similar to a native speaker.

4. Speaks fairly fluently with only occasional hesitation, false starts and modification of intended utterance. Speech is only slightly slower than that of a native speaker.

3. Speaks more slowly than a native speaker due to hesitations and word finding delays.

2. A marked degree of hesitation due to word finding delays or inability to phrase utterances easily.

1. Speech is quite disfluent due to frequent and lengthy hesitations or false starts.

Accuracy

5. Errors are barely noticeable.

4. Errors are not unusual, but rarely major.

3. Manages most common forms, with occasional errors; major errors present.

2. Limited linguistic control; major errors frequent.

1. Clear lack of linguistic control even of basic forms.

Complexity

5. Confidently attempts a variety of verb forms (e.g., passives, modals, tense and aspect), even if the use is not always correct. Regularly takes risks grammatically in the service of expressing complex meaning. Routinely attempts the use of coordination and subordination to convey ideas that cannot be expressed in a single clause, even if the result is occasionally awkward or incorrect.

4. Confidently attempts a variety of verb forms (e.g., passives, modals, tense and aspect), even if the use is not always correct. Takes risks grammatically in the service of expressing complex meaning. Regularly attempts the use of coordination and subordination to convey ideas that cannot be expressed in a single clause, even if the result is occasionally awkward or incorrect.

3. Mostly relies on simple verb forms, with some attempts to use a greater variety of forms (e.g. passives, modals, more varied tense and aspect). Some attempt to use coordination and subordination to convey ideas that cannot be expressed in a single clause.

2. Produces numerous sentence fragments in a predictable set of simple clause structures. If coordination and/or subordination are attempted to express more complex clause relations, this is hesitant and done with difficulty.

1. Produces mostly sentence fragments and simple phrases. Little attempt to use any grammatical means to connect ideas across clauses.

Appendix 2. Rater matrix.

	Students	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-77
Rater 1		X	X	X	X	X	X	X	X
Rater 2						X	X	X	X
Rater 3						X	X	X	X
Rater 4						X	X	X	X
Rater 5						X	X	X	X
Rater 6						X	X	X	X
Rater 7		X	X	X	X	X			
Rater 8		X	X	X					
Rater 9		X	X	X	X				
Rater 10		X	X	X	X				
Rater 11		X	X	X	X				
Rater 12		X	X	X	X	X			
Rater 13		X	X	X					
Rater 14		X	X	X	X	X			
Rater 15		X	X						

Appendix 3. CAF and fluency scores descriptive statistics.

	Mean	STD	N
Hesitations	1.69	1.20	154
Phonation Time Ratio	80.13	19.64	154
Filled Hesitations	73.72	21.50	154
Mean Length of Utterance	3.55	1.65	154
Total Speaking Time	81.53	32.51	154
Speech Rate	68.83	18.00	154
Filled Pauses	66.21	22.34	154
Guiraud's Index	4.84	0.75	154
Clauses per AS unit	1.49	0.39	154
Mean number of errors	1.47	0.83	154
Self-corrected errors	9.98	14.31	154

References

An, J., Macaro, E., & Childs, A. (2021). Classroom interaction in EMI high schools: Do teachers who are native speakers of English make a difference? *System, 98*. 10.1016/j.system.2021.102482.

Baddeley, A. (2007). *Working memory, thought, and action*. Oxford University Press.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89–110. 10.1191/0265532203lt245oa.

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30*(2), 159–175. 10.1177/0265532212455394.

Browne, K., & Fulcher, G. (2017). Pronunciation and intelligibility in assessing spoken fluency. In T. Isaacs, & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 37–53). Multilingual Matters.

Bui, G., & Huang, Z. (2018). L2 fluency as influenced by content familiarity and planning: Performance, measurement, and pedagogy. *Language Teaching Research, 22*(1), 94–114. 10.1177/1362168816656650.

Council of Europe. (2001). *Common European Framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Chou, M. H. (2018). Speaking anxiety and strategy use for learning English as a foreign language in full and partial English-medium instruction contexts. *TESOL Quarterly, 52*(3), 611–633. 10.1002/tesq.455.

De Jong, N. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly, 15*(3), 237–254. 10.1080/15434303.2018.1477780.

Eckes, T. (2011). *Introduction to Many-Facet Rasch measurement: analyzing and evaluating rater-mediated assessments*. Peter Lang.

Eckes, T., & Jin, K. Y. (2022). Detecting illusory halo effects in Rater-Mediated assessment: A mixture rasch facets modeling approach. *Psychological Test and Assessment Modeling, 64*(1), 87–111.

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2019). *Task-based language teaching: theory and practice*. Cambridge University Press.

Elder, C., & Iwashita, N. (2005). Planning for test performance: Does it make a difference?. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 219–239). John Benjamins.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 19*(4), 347–368. 10.1191/0265532202lt235oa.

Farrokhi, F., & Esfandiari, R. (2011). A Many-facet Rasch model to detect halo effect in three types of raters. *Theory and Practice in Language Studies, 1*(11), 1531–1540. 10.4304/tpls.1.11.1531-1540.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Studies in language testing 30 examining speaking* (pp. 65–112). Cambridge University Press.

Fisicaro, S. A., & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement, 14*(4), 419–429. 10.1177/014662169001400407.

Foster, P., & Skehan, P. (1996). The influence of planning time on performance in task-based learning. *Studies in Second Language Acquisition, 18*, 299–323.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*, 354–375. 10.1093/applin/21.3.354.

Fulcher, G. (2015). *Re-examining language testing a philosophical and social inquiry*. Routledge.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. Routledge.

Gilabert, R. (2007). The simultaneous manipulation of task complexity along planning time and (+/- here and now): Effects on oral production. In M. Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 44–68). Multilingual Matters.

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. 10.1177/0265532210364407.

Green, A. (2021). *Exploring language assessment and testing: Language in action* (2nd ed.). Routledge.

Guo, W., & Wind, S. A. (2021). An iterative parametric bootstrap approach to evaluating rater fit. *Applied Psychological Measurement, 45*(5), 315–330. 10.1177/01466216211013105.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency* (pp. 1–20). John Benjamins.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning, 51*, 401–436.

Isaacs, T., & Thomson, R. I. (2020). Reactions to second language speech: Influences of discrete speech characteristics, rater experience, and speaker first language background. *Journal of Second Language Pronunciation, 6*(3), 402–429. 10.1075/jslp.20018.isa.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed.), *Conversation analysis: studies from the first generation* (pp. 13–31). John Benjamins.

Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics, 39*(3), 569–591. 10.1017/S0142716417000534.

Kim, H. (2020). Effects of rating criteria order on the halo effect in L2 writing assessment: A many- facet Rasch measurement analysis. *Language Testing in Asia, 10*(16). 10.1186/s40468-020-00115-0.

Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing, 21*(1), 1–27. 10.1191/0265532204lt272oa.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26–43. 10.1016/j.asw.2007.04.001.

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35*(4), 477–499. 10.1177/0265532217710049.

Leung, C., & Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: Assessment as discourse and assessment of discourse. *Language Testing, 21*(3), 335–359. 10.1191/0265532204lt287oa.
Levelt, M. (1989). *Speaking: From intention to articulation*. MIT Press.
Li, L., Chen, J., & Sun, L. (2014). The effects of different lengths of pretask planning time on L2 learners' oral test performance. *TESOL Quarterly, 49*(1), 38–66.
Linacre, J. M. (2019). *Winsteps Rasch measurement computer program User's Guide*. Beaverton, Oregon: Winsteps.com.
Linacre, J. M. (n.d.). *Diagnosing Misfit*. Diagnosing misfit. Retrieved from https://winsteps.com/facetman/diagnosingmisfit.htm
Lo, Y., & Macaro, E. (2015). Getting used to content and language integrated learning: What can classroom interaction reveal? *The Language Learning Journal, 43*(3), 239–255. 10.1080/09571736.2015.1053281.
Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276. 10.1191/0265532202lt230oa.
McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal, 1*. 10.1017/S2041536210000012.
Muñoz, A., & Ramirez, M. (2015). Teachers' conceptions of motivation and motivating practices in second-language learning: A self-determination theory perspective. *Theory and Research in Education, 13*(2), 198–220. 10.1177/1477878515593885.
Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78*(2), 218–225. 10.1037/0021-9010.78.2.218.
Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.
Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189–227.
Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on oral task performance. *Language Testing, 31*(2), 147–175.
Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590–601. 10.1093/applin/amp045.
Pallotti, G. (2020). Measuring complexity, accuracy, and fluency (CAF). In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 201–210). Routledge.
Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards, & R. Schmidt (Eds.), *Language and communication* (pp. 191–225). Routledge.
Peltonen, P. (2017). Temporal fluency and problem-solving in interaction: An exploratory study of fluency resources in L2 dialogue. *System, 70*, 1–13. 10.1016/j.system.2017.08.009.
Pill, J., & Smart, C. (2020). Raters behaviour and training. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 135–144). Routledge.
Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal, 102*, 713–731. 10.1111/modl.12509.
Préfontaine, Y. (2013). Perceptions of French fluency in second language speech production. *The Canadian Modern Language Review, 3*, 324–348.
Qin, J., & Zhang, Y. (2022). Pre-task planning and discourse cohesion: Analysis of Chinese EFL learners' referential use in oral narratives. *Language Teaching Research, 26*, 60–78. 10.1177/1362168819883896.
Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes, 14*, 423–441. 10.1080/01638539109544795.
Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*, 465–493. 10.1177/0265532208094273.
Segalowitz, N. (2010). *Cognitive Bases of second language fluency*. Routledge.
Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18*, 373–391. 10.1177/026553220101800404.
Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics, 30*, 510–532. 10.1093/applin/amp047.
Skehan, P., & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.), *Planning and task-performance in a second language* (pp. 193–219). John Benjamins.
Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics in Language Teaching, 54*(2), 97–111. 10.1515/iral-2016-9992.
Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition, 42*(1), 143–167. 10.1017/S0272263119000421.
Suzuki, Y. (2021). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. *Language Learning, 71*(2), 285–325. 10.1111/lang.12433.
Suzuki, S., Kormos, & Uchihara, T. (2021). The Relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal, 105*(2), 435–463. 10.1111/modl.12706.
Tavakoli, P., & Hunter, A. M. (2018). Is fluency being 'neglected' in the classroom? Teacher understanding of fluency and related classroom practices. *Language Teaching Research*, 22(3), 330–349. 10.1.1177/1362168817708462
Tavakoli, P., Nakatsuhara, F., & Hunter, A. (2020). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal, 104*(1), 169–191. 10.1111/modl.12620.
Tavakoli, P., & Wright, C. (2020). *Second Language speech fluency: From research to practice*. Cambridge University Press. 10.1017/9781108589109.
Turner, C., & Upshur, J. (1996). Developing rating scales for the assessment of second language performance. *Australian Review of Applied Linguistics, 13*, 55–79.
VanVoorhis, C., & Morgan, B. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology, 3*, 43–50. 10.20982/tqmp.03.2.p043.
Xiao, Y., & Yang, M. (2019). Formative assessment and self-regulated learning: How formative assessment supports students' self-regulation in English language learning. *System, 81*, 39–49. 10.1016/j.system.2019.01.004.
Yorozuya, R., & Oller, J. W., Jr. (1980). Oral proficiency scales: Construct validity and the halo effect. *Language Learning, 30*, 135–153. 10.1111/j.1467-1770.1980.tb00155.x.
Zechner, K., & Evanini, K. (2020). *Automated speaking assessment using language technologies to score spontaneous speech*. Routledge.
Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31–50. 10.1177/0265532209360671.