Routledge
Taylor & Francis Group

Language Assessment Quarterly

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hlaq20

What Do the Measures of Utterance Fluency Employed in Automatic Speech Evaluation (ASE) Tell Us About Oral Proficiency?

Zoe L. Handley & Haiping Wang

To cite this article: Zoe L. Handley & Haiping Wang (08 Dec 2023): What Do the Measures of Utterance Fluency Employed in Automatic Speech Evaluation (ASE) Tell Us About Oral Proficiency?, Language Assessment Quarterly, DOI: 10.1080/15434303.2023.2283839

To link to this article: https://doi.org/10.1080/15434303.2023.2283839

Published online: 08 Dec 2023.

LANGUAGE ASSESSMENT QUARTERLY
https://doi.org/10.1080/15434303.2023.2283839

Routledge
Taylor & Francis Group

ARTICLE

OPEN ACCESS Check for updates

What Do the Measures of Utterance Fluency Employed in Automatic Speech Evaluation (ASE) Tell Us About Oral Proficiency?

Zoe L. Handley [a] and Haiping Wang [b]

[a]Department of Education, The University of York, York, UK; [b]East China University of Political Science and Law, Shanghai, People's Republic of China

ABSTRACT

This paper explores what the measures of utterance fluency typically employed in Automatic Speech Evaluation (ASE), i.e. automated speaking assessments, tell us about oral proficiency. 60 Chinese learners of English completed the second part of the speaking section of IELTS and six tasks designed to measure the linguistic knowledge and processing assumed to underpin second language speech production. A sample of eight native speakers rated the learners' oral productions for functional adequacy. Analyses of the data confirm: (1) articulation rate, mid-clause pause frequency, and repetition frequency predict functional adequacy, (2) breadth of lexical knowledge is the main predictor of articulation rate as well as functional adequacy, (3) speed of syntactic processing predicts end-clause pause duration and speed of lexical processing predicts mid-clause pause duration, and (4) measures of utterance fluency together account for 60% of the variation in functional adequacy scores. These findings suggest that articulation rate best reflects overall functional adequacy. Moreover, other measures of utterance fluency reflect different areas of underlying knowledge and processing, opening up the possibility of automating diagnostic speaking tests.

INTRODUCTION

As a result of recent advances in artificial intelligence and natural language processing, automated assessment is becoming a reality (Alexander et al., 2019), and there is growing interest in the development of summative and formative automated speaking tests. Examples of summative automated speaking tests include the *Pearson PTE Academic*, which is based on the *Versant* scoring model (Bernstein & Cheng, 2008), and *Educational Testing Service's (ETS) Test of English as a Foreign Language internet-Based Test (TOEFL iBT)*, which is based on the *SpeechRater* model (Zechner et al., 2009), among others (see Chen et al., 2018 for a review). Examples of software which provide formative feedback on speaking include *ELSA* Speak (www.elsaspeak.com) and Speechace (www.speechace.com). The potential these tests offer to reduce costs and increase the reliability of testing is highly

CONTACT Zoe L. Handley zoe.handley@york.ac.uk Department of Education, The University of York, York YO10 5DD, UK

attractive (Ginther et al., 2010; Wang et al., 2018). It is, however, necessary to critically examine these tests because they are typically based on stochastically derived scoring models which rely on a limited range of 'linguistic' features – typically a range of measures of speed (i.e. speech rate) and breakdown fluency (i.e. pausing behavior) along with a measure of accentedness (goodness of pronunciation), a measure of lexical diversity and a measure of grammatical accuracy. It is therefore important to understand the extent to which measures of utterance fluency (temporal features of utterances including speed, pausing and repair) reflect overall oral proficiency (the ability to use the spoken language to communicate effectively in real-world situations), and what they represent in terms of the linguistic knowledge and processes that underpin speech production. While a number of studies have attempted to establish the concurrent (e.g. Bridgeman et al., 2012), predictive (e.g. Harsch et al., 2017) and consequential validity (e.g. Farnsworth, 2013; Xi, 2007) of automated speaking tests such as *TOEFL iBT*, few studies have explored the construct validity (Kane, 2006; Williamson et al., 2012). Notable exceptions include Hsieh et al. (2019), Wang and Evanini (2019) and Yoon et al. (2019). These studies are, however, restricted to judgmental evaluations of the fit between the features upon which Automatic Speech Evaluation (ASE) models are based and the construct of interest, i.e. speaking quality and, few if any, empirical evaluations of the construct validity of ASE models have been conducted (Kane, 2006; Williamson et al., 2012).

Of the aforementioned 'linguistic' features, measures of utterance fluency have long featured in ASE scoring models (Bernstein et al., 2000) and continue to receive a high weighting compared to other measures (e.g. Higgins et al., 2011; Zechner et al., 2009). Concerns have, however, been raised about the validity of relying on measures of utterance fluency. First, native speakers vary considerably in their speech rate and their patterns of pausing and hesitation (De Jong, 2018). Second, learners report that they speak as quickly as possible without paying attention to meaning when using such systems (Xi et al., 2016). This paper contributes to the development of validity arguments for automated speaking tests (Bernstein et al., 2010) by exploring the concurrent and construct validity of measures of utterance fluency compared with measures of oral proficiency. This is achieved through a conceptual replication of Ginther et al. (2010) exploration of the relationship between measures of utterance fluency and scores on *Purdue's Oral English Proficiency Test* (*OEPT*) and de Jong et al.'s explorations of the componential structure of oral proficiency (De Jong et al., 2012) and oral fluency (De Jong et al., 2013). A conceptual replication is a study which attempts to verify the findings of a previous piece of research using a different research design or instruments to the original study (Porte, 2012). With the aim of establishing whether the findings of Ginther et al. (2010) and De Jong et al (2012, 2013). generalize to other language pairs, the replication reported here focuses on Chinese learners of English and is novel in its use of productive measures of the linguistic knowledge and processing that underpin speech production, including tasks developed for the specific purposes of this study.

LITERATURE REVIEW

Automatic speech evaluation (ASE)

ASE involves three main processes: (1) speech recognition, (2) feature extraction, and (3) grading. First, the speech recognizer generates a transcription of the speech produced by

the learner. This transcription is aligned with the audio and annotated with scores indicating how confident the recognizer is that a particular phone, word, or utterance has been correctly recognized as well as the results of lexical and syntactic analyses. From this transcription, a range of features of the learners' speech are then extracted. These include measures of utterance fluency, such as speech rate and average pause duration, measures of goodness of pronunciation, such as phone likelihood, and measures of the lexical and grammatical profile of the learners' productions, including lexical diversity and grammatical accuracy (Chen et al., 2018; Higgins et al., 2011; Wang et al., 2018; Zechner et al., 2009).

The grader or scoring model is a key component of the system and is developed by training an algorithm to predict human ratings from the acoustic and linguistic features extracted from learner productions on a corpus of rated learner productions (Wang et al., 2018; Zechner et al., 2009). Measures of fluency have long featured in such scoring models (Bernstein et al., 2000) and continue to be a core component (Chen et al., 2018; Ginther et al., 2010; Wang et al., 2018; Zechner et al., 2009). A wide range of different measures of utterance fluency have been considered in the development of such systems, including novel measures not seen in the wider Second Language Acquisition (SLA) literature in which fluency is measured (e.g. the task-based language teaching and study abroad literature). These include the standard deviation of silent pause duration and the standard deviation of chunk length among others (see Table 1). It's worth noting that none of the specific measures of utterance fluency have been adopted by all three models (see Table 1). Therefore, building validity arguments for objective fluency measures and their relationship with oral proficiency merits further consideration.

Oral proficiency (functional adequacy)

The measures of oral proficiency adopted by international standards (e.g. Magnan, 1988 cited in Iwashita et al., 2008) and popular language examinations (e.g. Iwashita et al., 2008) are typically based on asking human raters to provide a holistic score as well as the scores for a number of sub-components of proficiency, including among others vocabulary/expression, grammar, pronunciation/accent, accuracy, fluency, appropriateness, and comprehension/intelligibility (see Iwashita et al., 2008). Those working on SLA, on the other hand, tend to operationalize oral proficiency through CAF (Kuiken & Vedder, 2018), i.e. Complexity (elaborated language), Accuracy (error-free language), and Fluency (language produced in real-time; Ellis & Barkhuizen, 2005). Prioritizing an L1 point of reference over comprehensibility, it might be argued that both approaches focus on nativelikeness rather than the primary goal of most language learners, communicative or functional adequacy (De Jong et al., 2012; Révész et al., 2016), where communicative/functional adequacy refers to the ability to communicate successfully in real-world contexts. For this reason, some working in the field of language testing argue that assessment criteria ought to be based on insights from naïve speakers' ratings of the informativeness and comprehensibility of learner productions because linguists and language teachers "would find it almost impossible to ignore … errors in grammar, lexis and pronunciation" (De Jong et al., 2012, p. 15).

Table 1. Candidate measures of fluency for ASE scoring models.

Dimension of fluency		PTE Academic[1]	TOEFL iBT[2]	Cambridge[3]
Speed	Words per minute/second[4]		X	X
	Words per minute/second of speech time[5]	X	X	
	Mean phone(me) duration[6]	X		X
	Median phone(me) duration			X
	Standard deviation phone(me) duration			X
	Mean absolute deviation phone(me) duration			X
Breakdown	Total pause time	X		
	Filled pauses per minute/second		X	X
	Silent pauses per minute/second		X	
	Long[7] silent pauses per minute/second		X	X
	Silences per 100 words		X	
	Long silences per 100 words		X	
	Mean silent pause duration	X		X
	Median silent pause duration			X
	Standard deviation silent pause duration			X
	Mean absolute deviation silent pause duration			X
	Mean long silent pause duration			X
	Median long silent pause duration			X
	Standard deviation long silent pause duration			X
	Mean absolute deviation long silent pause duration			X
	Mean chunk length in words		X	
	Mean deviation chunk length in words		X	
	Mean chunk length in minutes/seconds		X	
	Mean deviation chunk length in seconds		X	
Repair	Disfluencies per minute/second			X
	Repetitions per 100 words		X	

[1]Versant model (Bernstein et al., 2011)
[2]SpeechRater model (Zechner et al., 2009)
[3]GP grader (van Dalen et al., 2015)
[4]I.e. speech rate measured in words, equivalent to mean word duration.
[5]I.e. articulate rate measured in words.
[6]Equivalent to articulation rate.
[7]A long silence refers to a silence that is greater than .5 seconds in duration (Zechner et al., 2009).

Oral fluency

As mentioned above, fluency as a component of oral proficiency refers to the speed and smoothness of learners' oral productions. Fluency thus defined has been explored from three different perspectives: (i) cognitive fluency, i.e. the efficiency of the processes thought to underpin oral proficiency, (ii) utterance fluency, i.e. the (temporal) features of utterances which are thought to reflect cognitive fluency, and (iii) perceived fluency, i.e. what inferences listeners make about cognitive fluency based on utterance fluency (Segalowitz, 2010).

ASE scoring models are normally based on measures of utterance fluency (Bernstein et al., 2011; Chen et al., 2018; Wang et al., 2018; Zechner et al., 2009). In the mainstream applied linguistics literature, these measures have been classified into measures of (1) speed fluency, i.e. rate of delivery, (2) breakdown fluency, i.e. the extent of interruptions, and (3) repair fluency, i.e. the number of self-corrections and repetitions (Segalowitz, 2010, p. 165). Further to this, Bernstein et al. (2011) on automatic speaking assessment distinguish structural fluency, e.g. clauses per minute and cohesives per minute, from phonological fluency, e.g. mean pause time and phones per minute. Of these various measures of utterance fluency, ASE scoring models normally rely on a selection of measures of speed fluency and breakdown fluency (see Table 1).

Broadly speaking, measures of utterance fluency are believed to reflect the ease and efficiency with which the processes which underpin speech production are functioning (Lennon, 1990), i.e. automaticity (Segalowitz, 2010). Speaking fluently, at a good pace without undue pausing and hesitation is crucial to keep listeners engaged and avoid communication breakdown (Rossiter, 2009). The relationship between linguistic processing and utterance fluency, however, remains little understood.

Validating the use of measures of utterance fluency as representations of oral proficiency

Research exploring the validity of using measures of utterance fluency as representations of oral proficiency can be classified into three types: (1) studies exploring the relationship between analytic ratings of specific features of oral productions (e.g. vocabulary, grammar, pronunciation and fluency) and global ratings of the same oral productions (see Koizumi et al., 2022 for a review), (2) studies exploring the relationship between objective measures of specific features of oral productions and global ratings of the same oral productions (e.g. De Jong et al., 2012), and (3) studies exploring the relationship between utterance fluency and difficulties in speech processing using introspective methods such as stimulated recall (e.g. Kahng, 2014). While some of these studies provide support for the use of measures of utterance fluency as reasonable representations of oral proficiency, with studies such as Iwashita et al. (2008) and Hulstijn et al. (2012) observing that measures of utterance fluency discriminate between different levels of proficiency and studies such as McNamara (1990) observing that ratings of fluency explain some of the variance in global ratings of proficiency, others do not (see De Jong et al., 2012). Moreover, it has been argued that studies which explore the relationship between analytic ratings and holistic ratings of oral proficiency based on the same learner productions are subject to circularity – "If one instructs raters to pay attention to speech rate and pausing, it is likely that the resulting ratings will be related to the objective measures speech rate and pausing" (De Jong et al., 2013, p. 896).

One project which has attempted to overcome the circularity inherent in much previous work is the *What is Speaking Project* (*WISP*; Hulstijn & Schoonen, 2004). This project focused on the development and validation of a Dutch speaking test and associated assessment rubrics. There were two phases to the validation of this speaking test:

(a) Exploration of the componential structure of speaking
(b) Assessment of the discriminant validity of the hypothesized components of speaking.

The circularity inherent in assessment validation work was overcome by collecting independent samples of learners' oral proficiency and their linguistic knowledge and processing and asking native speakers to rate learners' functional adequacy, i.e. the informativeness and comprehensibility of their speech (De Jong et al., 2012, 2013; Hulstijn et al., 2012). Reflecting models of speech production (Kormos, 2006; Levelt, 1989, 1999; Segalowitz, 2010), independent measures of the following dimensions of linguistic knowledge and processing were obtained: productive vocabulary knowledge, lexical retrieval speed, productive grammar knowledge, speed of sentence building, pronunciation skills, and speed of articulation.

With respect to the componential structure of speaking, De Jong et al. (2012) observed a strong relationship between all dimensions of linguistic knowledge and processing

measured and functional adequacy with the exception of speed of articulation, i.e. pronunciation. Together these measures accounted for 75% of variation in functional adequacy, with intonation (β = .34) and knowledge of vocabulary (β = .31) the strongest predictors. Similar models were obtained for learners classified as higher and lower proficiency. Increases in linguistic knowledge and processing, however, were found to have a greater impact on functional adequacy for higher proficiency learners than for lower proficiency learners.

Regarding the focus of this paper, understanding what the measures of utterance fluency employed in ASE scoring models represent, De Jong et al. (2013) also report a study exploring the relationship between linguistic knowledge and processing and utterance fluency in the context of developing rubrics for use in face-to-face language testing by human raters. This study found that linguistic knowledge and processing accounted for between 5% (mean pause duration) and 50% (mean syllable duration, i.e. articulation rate) of the variation in learners' levels of oral fluency, with linguistic knowledge and processing accounting for 22% of the variation in number of silent pauses, suggesting that articulation rate better reflects the componential structure of speaking than mean pause duration and pause rate. Correlations with articulation rate were strongest (i.e. > .50) for vocabulary knowledge, sentence building, i.e. morpho-syntactic encoding, and pronunciation quality, suggesting that articulation rate reflects vocabulary knowledge, syntactic processing, and pronunciation accuracy. Similar, but weaker, correlations with number of silent pauses were observed. A slightly different pattern of correlations was observed in Kahng's (2020) replication of De Jong et al. (2013) with Chinese learners of English where lexical retrieval speed and syntactic encoding speed most strongly correlated with mean syllable duration, i.e. articulation rate. Differentiating between mid- and final-clause pauses as well as filled versus silent pauses, Kahng (2020) also observed that different measures of linguistic knowledge and processing were more strongly related to some measures of utterance fluency than others. For example, vocabulary knowledge was also correlated with the number of mid-clause silent pauses, with syntactic encoding speed more strongly correlated with the number of mid-clause silent pauses than vocabulary knowledge and lexical retrieval speed. Phrase vocabulary knowledge and lexical retrieval speed, on the other hand, were more strongly correlated with the number of mid-clause filled pauses.

Reflecting De Jong et al. (2013) insights into the componential structure of fluency, Ginther et al. (2010) observed that articulation rate ($r = .61$, $p < .01$) was much more strongly related to scores on the *OEPT* than mean pause duration ($r = -.34$, $p < .01$). Ginther et al. (2010), however, observed a stronger relationship between mean syllables per run, their measure of pause rate, and scores on the *OEPT* ($r = .72$, $p < .01$) than might be expected given De Jong et al. (2013) findings in relation to the componential structure of measures of fluency.

The *WISP* project has made a significant contribution to the speech assessment literature by introducing an approach that overcomes the circularity of much work designed to understand the construct of speaking proficiency. It might, however, be considered that De Jong et al. (2013) sentence completion task is not sufficiently productive in nature and does not differentiate syntactic from morpho-syntactic processing. The same is true of similar tasks in Kahng's (2020) study.

In conclusion, a number of studies have explored the extent to which fluency reflects oral proficiency, i.e. concurrent and construct validity. The majority of these studies, however, might be considered to make circular arguments (De Jong et al., 2013). The remaining studies are also limited by a focus on exploring the relationship between receptive rather than productive linguistic knowledge and processing. Another limitation is the range of language pairs and proficiency levels explored. There is therefore a need for replications which focus on productive linguistic knowledge and processing and other language pairs.

Research questions

To address these issues, this paper replicates Ginther et al. (2010) and the *WISP* studies (De Jong et al., 2012, 2013), which explore the following questions, respectively:

(1) How does utterance fluency relate to functional adequacy?
(2) To what extent do measures of linguistic knowledge and processing predict functional adequacy?
(3) To what extent do measures of linguistic knowledge and processing predict utterance fluency?

Addressing the limitations noted at the end of the literature review, productive tasks are used to explore linguistic knowledge and processing and their relationship to fluency and oral proficiency among a group of Chinese learners of English.

As a replication of the aforementioned studies, it is anticipated that, of the measures of utterance fluency investigated, articulation rate will be most strongly related to functional adequacy (Ginther et al., 2010) and that vocabulary knowledge will be a strong predictor of functional adequacy (De Jong et al., 2012) as well as articulation rate (De Jong et al., 2013).

METHODOLOGY

This paper is based on the baseline data (i.e. Time 1) collected in a larger study designed to measure the impact of study abroad on oral fluency development among Chinese learners of English (see Handley & Wang, 2018). In this study, learners were asked to complete a speaking task and six tasks designed to tap the lexical and grammatical knowledge and processing involved in speech production as outlined by Levelt (1999). The time 1 data upon which this paper is based was collected shortly after the learners had started their master's program. The time 2 data was collected six months later. All tasks were administered in a single session, in the order in which they are presented below. Further to this, a sample of native speakers was recruited to rate the learners' productions in the speaking task.

Participants

Ethical approval was obtained from the principal researcher's departmental ethics committee. Participants were given £5 per hour compensation.

Chinese learners

Seventy-six Chinese learners of English studying for a master's volunteered to participate in the study. The learners were recruited from a range of programs including language and non-language majors in the UK and China. Data from 15 learners were omitted from the analyses presented here, two because they did not complete the second phase of the study, two due to missing functional adequacy ratings, and eleven because their performances were either too noisy to generate measures of fluency or because their scores were outliers on at least one of the measures of fluency due to background noise interfering with the automatic calculation of fluency measures. A further participant was omitted because data was missing for the speed of morphosyntactic processing task.

The final sample of 60 Chinese learners of English comprised 9 males and 51 females. Their ages ranged from 20 to 28 years (mean 23.02), and the age at which they began learning English ranged from 5 to 15 years old (mean 9.98). Based on the test scores they provided, their CEFR level was estimated to be between B1 to B2.

Native listeners of English

A further 8 native speakers of English volunteered to rate the learner's productions in terms of functional adequacy. All were university students. No training was provided beyond the task instructions.

Measures

Utterance fluency

Materials. Oral fluency in English was measured using the second part of the speaking section of IELTS, where test-takers are given a prompt card and asked to talk about a particular topic for two minutes (Case, 2008). As discussed above, the data reported here form part of a larger study designed to assess the impact of study abroad on oral fluency development. In order to eliminate potential learning effects from Time 1 to Time 2 in that study, two similar versions of the speaking task were used. In version A, learners were asked to talk about somewhere they had been on holiday. In version B, learners were asked to describe a journey they had been on (Case, 2008). Consequently, in this paper, half of the learners in each context completed the 'holiday' task, while the other half completed the 'journey' task.

Procedure. The tasks were presented on paper, and learners were given two minutes to plan before talking on the topic for two minutes. Recordings were made using a USB headset (Microsoft LifeChat LX-300) and the laptop's Sound Recorder.

Scoring. Coding proceeded as follows. First, the recordings were cleaned. That is, any off-task talk, coughs, and throat clearings were removed from the recordings (Burchfield & Bradlow, 2014). Then, the first 60 seconds of talk was extracted from each of the recordings. Then the recordings were coded automatically for speed and breakdown fluency using De Jong and Wempe's (2009) PRAAT script before hand-coding for pause location and repairs (see Table 2 for the full range of measures that were calculated and Foster & Skehan, 1996 for definitions of repairs including false starts, repetitions, and hesitations).

Table 2. Measures of oral fluency calculated in this study and their definitions.

Dimension of fluency	Measure
Proficiency/holistic	Phonation time ratio[1]
Speed fluency	Speech rate (syllables/total time)[2]
	Articulation rate (syllables/phonation time)
Breakdown fluency	Silent pauses per minute
	Long silent pauses per minute
	Silences per 100 syllables
	Long silences per 100 syllables
	Mean silent pause duration
	Standard deviation silent pause duration
	Mean absolute deviation silent pause duration
	Mean long (>500 ms) silent pause duration
	Standard deviation long (>500 ms) silent pause duration
	Mean absolute deviation long (>500 ms) silent pause duration
	Mean length of run in seconds
	Standard deviation length of run in seconds
	Mean absolute deviation length of run in seconds
	Mean length of run in syllables
	Standard deviation length of run in syllables
	Mean absolute deviation length of run in syllables
Repair fluency	Repetitions
	Lexical substitutions
	False starts

[1]While not technically a measure of fluency in the narrow sense of fluidity (Segalowitz et al., 2017), phonation time ratio is included as a measure of fluency in the ASE literature and was included in our preliminary analyses because it is a standardized measure of the volume of speech produced and hence considered reflective of fluency in the broad sense of proficiency (Segalowitz et al., 2017).

[2]De Jong and Wempe's (2009) PRAAT script is based on syllable nuclei. Duration of syllables is not the same as duration between syllable nuclei. It is therefore not possible to calculate the standard deviation and mean absolute deviation in syllable duration. For this reason, speed fluency is conceptualized as speech rate and articulation rate in this study, rather than as Mean Syllable Duration (MSD).

Following Zechner et al. (2009), bivariate Pearson correlations were run between all measures of oral fluency. High inter-correlations suggest that measures are tapping the same underlying dimension of oral fluency and it is redundant to include both in an ASE algorithm. Only one measure was therefore retained for each such correlation identified in our data. These were selected on the basis of their 1) independence of other dimensions of fluency, 2) congruence with the defined underlying construct, and 3) interpretability. In line with (Segalowitz 2017) analysis of the more narrow range of measures typically reported in the SLA literature, the range of fluency measures explored in the analyses presented in this paper was therefore reduced to: articulation rate, unfilled pauses per 100 syllables and mean unfilled pause duration, including separate measures for unfilled pauses occurring at the end of and mid-clause, and repetitions per 100 syllables (see De Jong, 2016; Tavakoli, 2011).[1]

Reliability. Reliability was not relevant because coding was automated.

Comparability. Independent samples t-tests confirmed no significant differences between learners' performance on the two different versions of the speaking task in terms of measures of utterance fluency (Articulation rate: $t(58) = .071$, $p = .94$; Mean length of run: $t(58) = .533$, $p = .60$; Mean pause duration: $t(58) = .125$, $p = .90$).

[1]Independent clauses, sub-clausal units, and subordinate clauses were all considered to be clauses and defined following Foster et al. (2000).

Functional adequacy

Materials. An adapted translation of De Jong et al. (2012) functional adequacy scale was used in this study. The experimental stimuli comprised the learners' productions in the English speaking task. The practice stimuli comprised data collected from learners who did not participate in both phases of the study.

Procedure. The eight English listeners who were recruited to rate the learners' productions were randomly allocated to one of four groups. Each rater only rated each participant once. That is, they only rated either the participants' Time 1 production or Time 2 production. Each rater also only rated half of the participants studying in the UK and half of the participants studying in China. To mitigate order effects, Time, Country, and order of presentation were counterbalanced across the four groups of raters. Having completed the practice trials, the experimental trials were presented in randomized order.

Reliability. A good level of inter-rater reliability was achieved (ICC (average measures) = .85 on average for Time 1). Ratings of the functional adequacy of the learner productions were therefore averaged across the four listeners that rated each learner production.

Comparability. Independent samples t-tests confirmed no significant differences between learners' performance on the two different versions of the speaking task in terms of either functional adequacy ($t(58) = -.296$, $p = .77$).

Breadth of vocabulary knowledge

Materials. Breadth of vocabulary knowledge was assessed using a subset of items from Laufer and Nation's (1999) *Productive Levels Test*, comprising the 18 items at the first 2000 word families level and 18 items from the *University Word List* (Xue & Nation, 1984). The 2000 word level was combined with the university word list level, because developing fluency involves "getting good at using what is already known" (Nation, 2007, p. 8). The data reported here form part of a larger study in which two versions of the *Productive Levels Test* were counterbalanced (see above). Half of the learners therefore completed version A, while the other half completed version B.

Procedure. Learners were given 15 minutes to complete the task, which was administered on paper.

Scoring. One point was awarded for each correct response, and a point was only awarded if the item was spelled correctly.

Reliability . Both versions of the test had high reliability (Cronbach's $\alpha = .82$and .81, respectively).

Comparability. An independent-samples t-test confirmed no significant differences between learners' performance on the two versions of the test ($t(58) = 1.067$, $p = .29$).

Depth of vocabulary knowledge

Materials. Depth of vocabulary knowledge was assessed using a subset of Read's (1993) *Word Associates Test* comprising 10 items which only included words at the first 2000 word families level – fluency, as highlighted above, is about using known language efficiently (Nation, 2007). These items were: beautiful, bright, calm, natural, fresh, general, bare, conscious, convenient, and curious.

Procedure. Learners were given 10 minutes to complete the task, which was administered on paper. For each item learners were tasked with identifying four associates, including at least one semantic and one syntagmatic associate, from a set of eight options (Read, 2004).

Scoring. One point was awarded for each correct response. If learners marked more than four responses, they received a mark of zero for that item.

Reliability. A high level of reliability was confirmed for the test (Cronbach's $\alpha = .73$).

Comparability. Comparability was not relevant because only one version of this task was used.

Speed of lexical retrieval

Materials. Lexical retrieval time was obtained via a picture naming task. The experimental stimuli comprised 16 words from Farrell and Abrams (2014) and the corresponding images from Brodeur et al (2010, 2014). *Bank of Standardized Stimuli*. The 16 selected words were those which were found to be known by most learners in Handley and Wang (2018) which employed an earlier version of this test to explore similar questions in the same population (see Appendix A for a full list of stimuli).

Procedure. The task was presented using *PsychoPy v.1.8* (Peirce, 2007). On each trial, the learners were presented a picture and asked to name it as quickly as possible. More specifically, on each trial the learners were instructed to press the spacebar when ready. Having pressed the spacebar a fixation point appeared for 500 ms. This was followed by presentation of a target picture which remained on screen for 5000 ms. The learners' responses were recorded using a USB headset (*Microsoft LifeChat LX-300*). The experimental stimuli were presented to each learner in a different random order. Before starting the task in earnest, the learners completed a block of four practice trials and received feedback from the researcher.

Scoring. One point was awarded for each correct response. Reaction and completion times were coded automatically for correct responses only using a *PRAAT* script developed for the specific purposes of this study, then checked by hand by a research assistant.

Reliability. A moderate level of reliability was confirmed for the test (Cronbach's $\alpha = .64$).

Comparability. Comparability was not relevant because only one version of this task was used.

Grammar knowledge

Materials. Grammar knowledge was assessed using Norris (2005) *Grammar Ability Finder*, an instrument designed to tap learners' ability to apply five word order rules. The data reported here form part of a larger study in which two versions of the *Grammar Ability Finder* were counterbalanced (see above). Half of the learners therefore completed version A, while the other half completed version B.

Procedure. Learners were given 10 minutes to complete the task, which was administered on paper.

Scoring. One point was awarded for each correct response. Responses in which some, but not all of the words were in the correct order, did not receive a point. Spelling mistakes were, however, ignored to avoid measuring vocabulary knowledge in the grammar test.

Reliability. Both versions of the test had relatively low levels of reliability (Cronbach's $\alpha = .44$ and .55, respectively) – these lower levels of reliability might be explained by the fact that a number of items did not discriminate between our participants.

Comparability. An independent-samples t-test confirmed no significant differences between learners' performance on the two versions of the test ($t(58) = -.798$, $p = .43$).

Speed of morphosyntactic processing

Materials. A modified version of Engelhardt et al. (2012) sentence construction task was used to measure learners' morphosyntactic processing. Designed to tap learners' processing of number and tense agreement, the 16 experimental stimuli which were based on words in the *General Service List* (West, 1953) comprised a pronoun, noun, or two plus a noun, and an infinitive verb (e.g. "two + shop" and "to close"; see Appendix B, Table B1, for a full list of stimuli).

Procedure. The task was presented using *PsychoPy v.1.8* (Peirce, 2007). On each trial, learners were presented two noun phrases and a verb phrase accompanied by a time phrase (i.e. "everyday", "now", "yesterday", or "tomorrow") indicating the tense of the utterance they were required to produce. Their task was to construct an utterance in the tense indicated using the constituents presented as quickly as possible. Altogether there were four quartiles of trials labeled with different time phrases respectively and with either singular or plural forms of nouns as the subject (see Figure 1 below for some examples).

More specifically, on each trial the learners were instructed to press the spacebar when ready. Having pressed the spacebar a fixation point appeared for 500 ms. This was followed by presentation of the target stimulus set which remained on screen for 10,000 ms. The learners' responses were recorded using a USB headset (*Microsoft LifeChat LX-300*). The experimental stimuli were presented to each learner in a different random order. Before

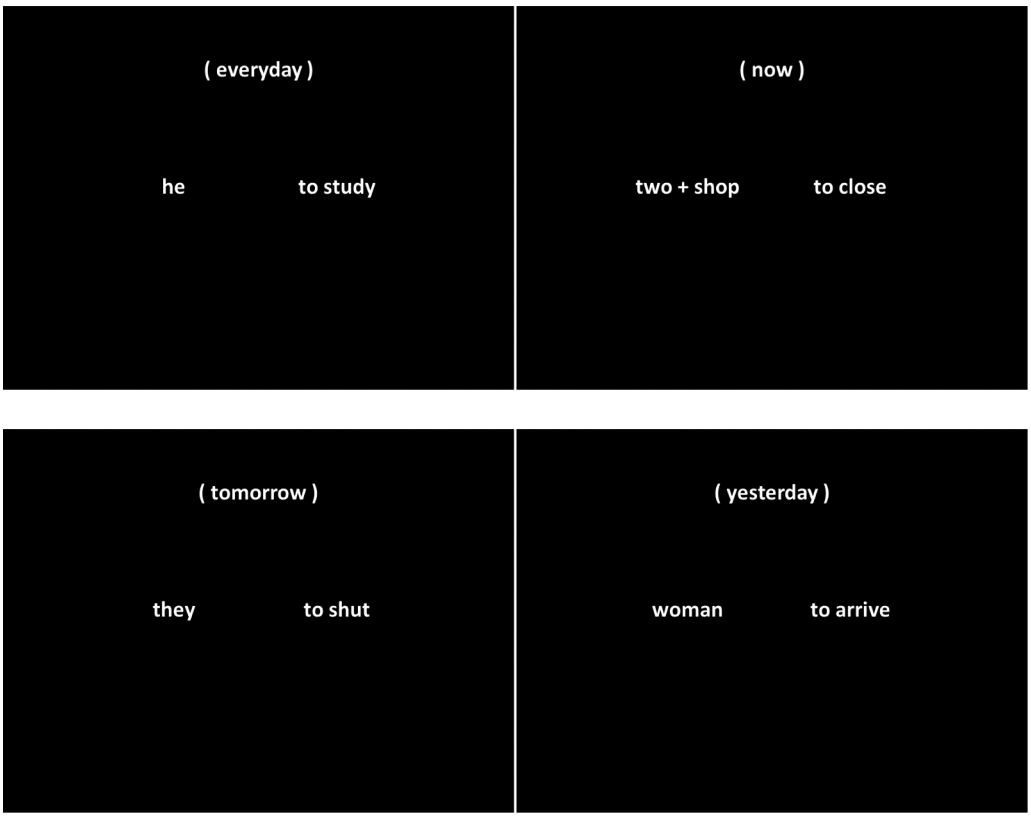

Figure 1. Screenshots illustrating the different conditions of the sentence inflection task.

starting the task in earnest, the learners completed a block of six practice trials and received feedback from the researcher (see Appendix B for task instructions).

Scoring. This task was scored in a similar way to the grammar knowledge test. That is errors in pronunciation and conjugation were ignored. Reaction and completion times were coded automatically for correct responses only using a *PRAAT* script developed for the specific purposes of this study, then checked by hand by a research assistant.

Reliability. An adequate level of reliability was confirmed for the test (Cronbach's $\alpha = .77$).

Comparability. Comparability was not relevant because only one version of this task was used.

Speed of syntactic processing

Materials. A modified version of Engelhardt et al. (2012) sentence construction task was used to measure learners' syntactic processing. The experimental stimuli comprised 24 sets of two noun phrases, one referring to an animate object and one referring to an inanimate object, and a verb phrase (e.g. "the milk", "the cat" and "is drinking") based on words in the General Service List (GSL; West, 1953; see Appendix C, Table C1, for a full list of stimuli).

Procedure. The task was presented using *PsychoPy v.1.8* (Peirce, 2007). On each trial, learners were presented two noun phrases and a verb phrase accompanied by a punctuation mark indicating the type of utterance they were required to produce. Their task was to construct an utterance of the type indicated using the constituents presented as quickly as possible. On one half of the trials, sentence constituents appeared in the order animate object, inanimate object, verb; on the other half, the sentence constituents appeared in the order inanimate object, animate object, verb. On one third of the trials, a full stop was displayed at the top of the screen, indicating that the learners should produce a declarative statement; on another third, a question mark was displayed, indicating that learners should produce an interrogative; and, on the final third, a minus sign was displayed indicating that learners should produce a negative statement (see Figure 2 for some examples).

The remaining procedure was the same as for the morpho-syntactic processing (see Appendix C for task instructions).

Scoring. The task was scored and coded in the same way as the morpho-syntactic processing task (see above).

Reliability. An adequate level of reliability was confirmed for the test (Cronbach's $\alpha = .78$).

Comparability. Comparability was not relevant because only one version of this task was used.

Analysis

Before running descriptive statistics and inferential analyses, the data was checked for normality and outliers (see Participants).

Descriptive statistics were also generated for all of the measures and comparisons were made with previous studies to help understand the comparability of the data set.

Our main questions relating to what areas of linguistic knowledge and processing involved in speech production measures of utterance fluency represent were then explored through correlational analyses and linear regression. Research question 1, asking how utterance fluency relates to functional adequacy, was evaluated by creating

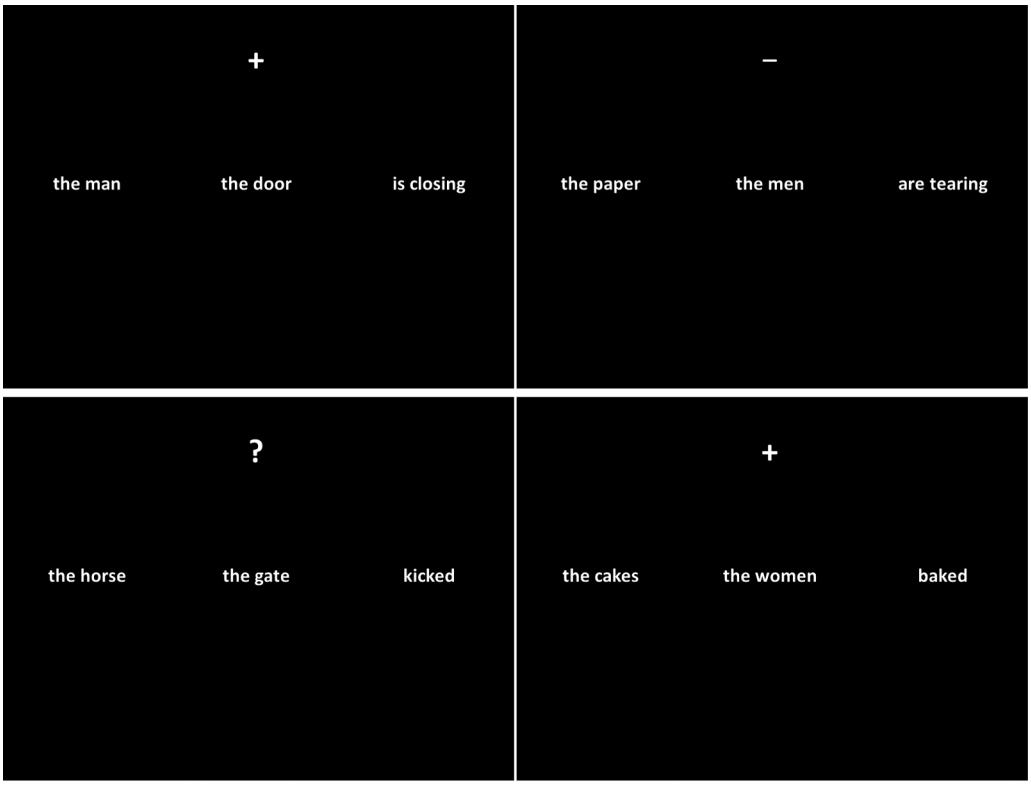

Figure 2. Screenshots illustrating the different conditions of the sentence transformation task.

a linear regression model with functional adequacy as the outcome and the six utterance fluency measures as predictor variables. This model can shed light on which measures of utterance fluency might be underpinned by a similar construct to functional adequacy. Research question 2, asking to what extent measures of linguistic knowledge and processing predict functional adequacy, was evaluated by creating a linear regression model with functional adequacy as the outcome and the six measures of linguistic knowledge and processing as predictor variables. This model can shed light on the componential structure of speaking proficiency. Research question 3, asking to what extent measures of linguistic knowledge and processing predict utterance fluency, was evaluated by creating a series of linear regression models, one per measure of utterance, with the measure of utterance fluency as the outcome and the six measures of linguistic knowledge and processing as predictor variables. These models can shed light on the extent to which measures of utterance fluency represent the same construct as functional adequacy. If a model of a measure of utterance fluency has a similar structure to the model of functional adequacy, it would suggest that that measure is underpinned by a similar construct and might be used alone as a proxy for functional adequacy. If a model of a measure of utterance fluency reflects some dimensions of functional adequacy, but not others, it would suggest that it under-represents the construct underpinning functional adequacy and would need to be supplemented with other measures in order to fully represent the construct. The analyses were carried out in R (R Core Team, 2015) using the tidy verse package (Wickham et al., 2019).

RESULTS AND DISCUSSION

Descriptive statistics

Table 3 shows the means and standard deviations for ratings of functional adequacy, measures of utterance fluency and measures of linguistic knowledge and processing. Note that a higher score represents better performance on measures of functional adequacy, articulation rate, breadth and depth of vocabulary knowledge, and grammar knowledge, whereas a lower score represents better performance on measures of pause frequency, mean pause duration, repetition frequency, and speed of lexical, syntactic and morphosyntactic processing. Taking this into consideration, this data suggests that the learners in this study have moderate levels of functional adequacy ($M = 3.90$; SD = 1.11) similar to those observed for the learners of Dutch in De Jong et al. (2012) study which employed a similar measure of oral proficiency. Levels of utterance fluency, however, differ, with the Chinese learners of English in this study uttering fewer syllables per second ($M = 3.26$; SD = 0.42) but making fewer repairs ($M = 3.29$; SD = 3.07), paused less often (end clause pause frequency: $M = 3.60$, SD = 2.10; mid-clause pause frequency: $M = 7.66$, SD − 4.13) and pausing for less time on average (end clause pause duration: $M = 519$, SD = 222; mid-clause pause duration: $M = 486$, SD = 156) than the learners of Dutch in De Jong et al (2013, 2015). and the Chinese learners of English in Kahng's (2020) study. The differences with De Jong et al (2013, 2015). studies might be explained by the transfer of speech rate and patterns of pausing associated with the learners L1 to their L2. While information rate, that is "the average rate at which information is emitted" (Coupé et al., 2019, p. 1), is somewhat universal across languages at 39 bits per second, structural differences between languages result in differences in the amount of information they encode per syllable and hence speech rate (see Coupé et al., 2019), and Mandarin has significantly fewer distinct syllables than English and other Indo-European languages and is associated with a slower speech rate than English (Coupé et al., 2019). The differences with Kahng's (2020) study might be explained by differences in the task – the learners' in Kahng's (2020) study completed two tasks from the *TOEFL iBT* which required them to express preferences and express and support their decisions, respectively. It should also be noted that Kahng's (2020) learners were of a wider

Table 3. Descriptive statistics (means and standard deviations) for measures of linguistic knowledge and processing ($N = 60$).

Measure	Mean	Standard deviation
Functional adequacy (0–7)	3.90	1.11
Articulation rate (syll/sec)	3.01	0.50
End clause pause frequency (per 100 syll)	3.60	2.10
Mid-clause pause frequency (per 100 syll)	7.66	4.13
End clause pause duration (ms)	519	222
Mid-clause pause duration (ms)	486	156
Repetition frequency (per 100 syll)	3.29	3.07
Breadth of vocabulary knowledge (max 36)	19.58	5.74
Depth of vocabulary knowledge (max 40)	30.97	4.39
Vocabulary knowledge (max 16)	11.90	2.49
Speed of lexical retrieval (ms)	3193	387
Grammar knowledge (max 12)	9.90	1.61
Syntactic knowledge (max 24)	17.40	4.26
Speed of syntactic processing (ms)	6277	721
Morphosyntactic knowledge (max 16)	11.18	3.38
Speed of morphosyntactic processing (ms)	5523	1225

range of proficiency levels (CEFR C1 to A2) than those in the present study whose proficiency range from B1 to B2. This is also a possible explanation for the lower variance in utterance fluency across learners compared to previous studies.

Speed of linguistic processing was considerably slower than reported in De Jong et al. (2013) previous work on learners of Dutch, Kahng's (2020) research on Chinese learners of English, and other work reporting similar measures (Speed of lexical processing $M = 3193$, SD = 387; Speed of syntactic processing $M = 6277$, SD = 721, Speed of morphosyntactic processing $M = 5523$, SD = 1225). The slower reaction times on the picture naming task might be explained by the fact that lexical/conceptual representations include image representations which are connected to the linguistic/cultural context in which they were acquired (Jared et al., 2013) and, as far as it is possible to establish, the Brodeur et al (2010, 2014). stimuli have not been normed for Mandarin unlike the stimuli in Kahng (2020). Alternatively, if the learners in this study are accessing the English items via their L1, i.e. translation (Jiang, 2002), the slower reactions times may reflect the fact that naming latencies for Mandarin tend to be longer than for other languages – Mandarin has a relatively small syllable inventory compared with other languages resulting in a larger number of polysyllabic words and greater competition during lexical selection (Bates et al., 2003; Weekes et al., 2007).

The slower response times on the syntactic and mopro-syntactic processing tasks might be explained by the fact that the tasks are more productive than De Jong et al. (2013) and Kahng's (2020) sentence completion tasks. Moreover, the differences with De Jong et al. (2013) study might be explained by cross-linguistic differences. Mandarin uses a different script to English; Mandarin, unlike English, does not form questions through Wh-movement (Huang, 1988); Mandarin verbs are only marked for aspect (Tardif et al., 1997); and, there is no subject-verb agreement in Mandarin (Huang, 1989; Jaeggli & Safir, 1989 cited in Tardif et al., 1997) and learners have been shown to be insensitive to it in English (Jiang, 2004).

These differences in linguistic background, proficiency levels and task highlight the importance of replicating previous research to confirm whether the findings generalize to other populations as we do in this paper.

The relationship between oral fluency and functional adequacy

To explore the question of whether utterance fluency reflects global proficiency, ratings of learners' functional adequacy in the IELTS-style speaking task were regressed on the measures of utterance fluency which were obtained from them. These results, which are summarized in Table 4, show that articulation rate ($\beta = 0.93$, 95% CI = [0.52, 1.35], $t = 4.51$, $p < .001$), mid-clause pause frequency ($\beta = -0.13$, 95% CI = [−0.19, −0.06], $t = -3.91$, $p < .001$) and repetition frequency ($\beta = -0.14$, 95% CI = [−0.21, −0.07], $t = -4.08$, $p < .001$) are predictors of functional adequacy. In other words, the faster the learners spoke and the less frequently they paused mid-clause and the less frequently they repeated themselves, the more proficient they were perceived to be. Together the measures of utterance fluency account for 60% of the variation in functional adequacy scores, which suggests that other measures (including quality measures) should be adopted to more adequately represent the construct of oral proficiency.

Table 4. Correlations between measures of utterance fluency and functional adequacy, and model results.

Variable	Correlation	Estimate (SE)	95% CI	*t*	*p*
Intercept		1.72 (0.75)	0.22–3.22	2.30	0.025
Articulation rate	.49**	0.93 (0.21)	0.52–1.35	4.51	<0.001
End clause pause frequency	.05	0.03 (0.05)	−0.08–0.13	0.47	0.641
Mid-clause pause frequency	−.49**	−0.13 (0.03)	−0.19 – −0.06	−3.91	<0.001
End clause pause duration	.08	0 (0.00)	−0.00–0.00	0.13	0.896
Mid-clause pause duration	−.20	0 (0.00)	−0.00–0.00	1.69	0.097
Repetition frequency	−.51**	−0.14 (0.03)	−0.21 – −0.07	−4.08	<0.001

**$p < .01$, $R^2 = .60$.

These findings reflect those of Yan (2020) who observed moderate to strong associations between common measures of utterance fluency including articulation rate and proficiency scores. The correlations are, however, weaker than those observed in Ginther et al. (2010) investigation of the relationship between measures of utterance fluency and scores on the *OEPT*. Such differences might be explained by the circularity inherent in validation work highlighted by De Jong et al. (2012) – the *OEPT* descriptors like those employed in other assessments make specific reference to fluency-related characteristics of oral productions (see Tavakoli et al., 2017).

The question therefore remains whether these measures of utterance fluency represent the same construct as functional adequacy.

The relationship between linguistic knowledge and processing and functional adequacy

To understand the construct of functional adequacy, its componential structure was explored by regressing ratings of learners' functional adequacy in the IELTS-style speaking task on measures of their linguistic knowledge and processing (see Table 5). In line with previous research exploring the componential structure of functional adequacy (e.g. De Jong et al., 2012), breadth of lexical knowledge ($\beta = 0.13$, 95% CI = [0.07, 0.18], $t = 4.52$, $p < .001$) was a strong predictor of functional adequacy. In other words, as learners' vocabulary size increases so does their overall proficiency.

That breadth of lexical knowledge predicts functional adequacy is unsurprising. The relationship between breadth of lexical knowledge and measures of reading, writing and listening proficiency is well-established (Milton, 2013). With respect to speaking, it received greater weighting than other areas of linguistic knowledge and processing in De Jong et al

Table 5. Correlations between measures of linguistic knowledge and processing and functional adequacy and model results.

	Correlations	Estimate (SE)	95% CI		*t*	*p*
Intercept		2.69(1.78)	−0.79	6.18	1.52	.135
Breadth of vocabulary knowledge	0.56**	0.13(0.03)	0.07	0.18	4.52	<0.001
Depth of vocabulary knowledge	0.04	−0.02(0.03)	−0.08	0.04	−0.62	.536
Grammar knowledge	0.10	−0.11 (0.10)	−0.30	0.08	−1.13	.262
Speed of lexical processing	0.22	0.00 (0.00)	0.00	0.00	1.44	.155
Speed of syntactic processing	−0.11	−0.00 (0.00)	0.00	0.00	−0.54	.589
Speed of morphosyntactic processing	−0.29*	−0.00 (0.00)	0.00	0.00	−0.69	.490

**$p < .01$, $R^2 = .37$.

(2012, 2013). research exploring the relationship between objective measures of linguistic knowledge and processing and ratings of functional adequacy as well as in studies exploring the relationship between analytic ratings of oral proficiency and global ratings (see De Jong et al., 2012). It is important to acknowledge that not all measures included in De Jong et al. (2012) study were included in the present one, including for example intonation. The absence of relationships between other dimensions of linguistic knowledge and processing and functional adequacy as observed in De Jong et al. (2012) might, however, be explained by the cross-linguistic differences between Mandarin and English highlighted above in the discussion of the reaction times in the syntactic and morpho-syntactic processing tasks above.

The relationship between linguistic knowledge and processing and utterance fluency

To understand the componential structure of the different measures of utterance fluency, each measure was regressed on measures of linguistic knowledge and processing in turn (see Table 6). The results for articulation rate are similar to those for functional adequacy, with breadth of lexical knowledge the only dimension of linguistic knowledge and processing predicting articulation rate ($\beta = 0.04$, 95% CI = [0.01, 0.06], $t = 2.86$, $p < .01$). In other words, the larger the learners' vocabulary, the more quickly they spoke. Speed of lexical processing, on the other hand, is the only significant predictor of end clause pause duration ($\beta = -0.19$, 95% CI = [−0.33, −0.02], $t = -2.33$, $p = .02$). In other words, learners who took less time to retrieve lexical items paused for longer at the end of clauses. Speed of syntactic processing is the only significant predictor of mid-clause pause duration ($\beta = 0.01$, 95% CI = [0.03, 0.16], $t = 3.06$, $p < .01$). In other words, learners who could more quickly transform sentences paused for less time mid-clause. As for repetitions per 100 syllables, none of the measures of linguistic knowledge and processing were found to be significant predictors of this measure of utterance fluency. The measures of linguistic knowledge and processing used in this study only accounted for between 6% and 23% variation in measures of utterance fluency, suggesting that they may represent non-linguistic factors.

That measures of vocabulary breadth, depth and processing are associated with measures of utterance fluency has been replicated across a number of studies, suggesting that language is more lexically than syntactically based (Van Moere, 2012). The precise nature of the relationship is, however, unclear with some studies observing a relationship between breadth of vocabulary knowledge and articulation rate (e.g. De Jong et al., 2013), while others do not (e.g. Kahng, 2020; see; Ebrahimi, 2021 for a comprehensive review). The relationship between grammar knowledge and processing and utterance fluency on the other hand is less well researched. A notable exception is De Jong et al. (2013) examination of the relationship between linguistic knowledge and processing and utterance fluency among advanced learners of Dutch. In contrast with this study, De Jong et al. (2013) observed a strong relationship between sentence building speed and number of silent pauses, but not between sentence building speed and mean silent pause duration. As discussed above there are important cross-linguistic differences between Mandarin and Dutch which might explain the longer processing speeds observed in this study compared to those observed in De Jong et al. (2013) and hence the differences in observed relationships between processing and utterance fluency. Together these results, while confirming that measures of utterance fluency reflect lexical knowledge and processing, emphasize the importance of carrying out research with a wider range of L1-L2 pairs to fully understand

Table 6. Correlations between measures of linguistic knowledge and processing and measures of utterance fluency, and model results for measures of utterance fluency.

Model	Correlations	Estimate (SE)	95% CI	*t*	*p*
Articulation Rate					
Intercept		5.21 (0.84)	3.53–6.90	6.21	<0.001
Breadth of vocabulary knowledge	0.36**	0.04 (0.01)	0.01–0.06	2.86	0.006
Depth of vocabulary knowledge	0.04	0 (0.01)	−0.03–0.03	−0.10	0.924
Grammar knowledge	0.04	−0.08 (0.05)	−0.18–0.01	−1.81	0.076
Speed of lexical processing	−.28*	0 (0.00)	−0.00–0.00	−1.48	0.144
Speed of syntactic processing	−0.35**	0 (0.00)	−0.00–0.00	−1.64	0.106
Speed of morphosyntactic processing	−0.32**	0 (0.00)	−0.00–0.00	−1.04	0.301
R^2	0.32				
End clause pause frequency					
Intercept		6.85 (4.12)	−1.42–15.12	0.10	0.102
Breadth of vocabulary knowledge	0.05	0.02 (0.06)	−0.11–0.15	0.75	0.748
Depth of vocabulary knowledge	0.13	0.06 (0.07)	−0.08–0.19	0.39	0.393
Grammar knowledge	0.00	−0.13 (0.23)	−0.58–0.33	0.58	0.579
Speed of lexical processing	−0.15	0 (0.00)	−0.00–0.00	0.47	0.465
Speed of syntactic processing	−0.18	0 (0.00)	−0.00–0.00	0.39	0.39
Speed of morphosyntactic processing	−0.04	0 (0.00)	−0.00–0.00	0.90	0.895
R^2	0.06				
Mid-clause pause frequency					
Intercept		12.69 (7.94)	−3.23–28.61	1.60	0.116
Breadth of vocabulary knowledge	−0.11	−0.12 (0.12)	−0.37–0.12	−1.01	0.318
Depth of vocabulary knowledge	0.11	0.15 (0.13)	−0.11–0.41	1.16	0.252
Grammar knowledge	−0.06	0.02 (0.44)	−0.86–0.89	0.04	0.97
Speed of lexical processing	−0.23	0 (0.00)	−0.01–0.00	−1.91	0.062
Speed of syntactic processing	−0.03	0 (0.00)	−0.00–0.00	0.41	0.685
Speed of morphosyntactic processing	0.05	0 (0.00)	−0.00–0.00	−0.19	0.847
R^2	0.10				
End clause pause duration					
Intercept		97.22 (397.45)	−699.97–894.41	0.25	0.808
Breadth of vocabulary knowledge	0.29*	3.96 (6.19)	−8.46–16.37	0.64	0.525
Depth of vocabulary knowledge	0.18	5.68 (6.55)	−7.47–18.82	0.87	0.39
Grammar knowledge	0.33**	37.6 (21.88)	−6.29–81.49	1.72	0.092
Speed of lexical processing	−0.26**	−0.18 (0.08)	−0.33 – −0.02	−2.33	0.024
Speed of syntactic processing	0.01	0.06 (0.05)	−0.03–0.15	1.39	0.171
Speed of morphosyntactic processing	0.01	−0.01 (0.03)	−0.06–0.05	−0.25	0.802
R^2	0.22				
Mid-clause pause duration					
Intercept		−283.8 (280.7)	−846.81–279.22	−1.01	0.317
Breadth of vocabulary knowledge	0.08	0.04 (4.37)	−8.73–8.81	0.01	0.992
Depth of vocabulary knowledge	0.16	6.81 (4.63)	−2.47–16.10	1.47	0.147
Grammar knowledge	0.13	15.46 (15.45)	−15.53–46.46	1.00	0.322
Speed of lexical processing	−0.03	−0.07 (0.05)	−0.17–0.04	−1.26	0.215
Speed of syntactic processing	0.34**	0.1 (0.03)	0.03–0.16	3.06	0.003
Speed of morphosyntactic processing	0.18	0 (0.02)	−0.04–0.04	0.01	0.991
R^2	0.21				
Repetition frequency					
Intercept		12.4 (5.75)	0.87–23.93	2.16	0.036
Breadth of vocabulary knowledge	−0.26*	−0.06 (0.09)	−0.24–0.12	−0.68	0.498
Depth of vocabulary knowledge	−0.06	0.04 (0.09)	−0.15–0.23	0.42	0.678
Grammar knowledge	−0.34**	−0.54 (0.32)	−1.18–0.09	−1.7	0.091
Speed of lexical processing	−0.14	0 (0.00)	−0.00–0.00	−1.16	0.253
Speed of syntactic processing	0.03	0 (0.00)	−0.00–0.00	0.03	0.973
Speed of morphosyntactic processing	0.07	0 (0.00)	−0.00–0.00	0.08	0.94
R^2	0.15				

**$p < .01$.

the nature of the constructs of oral fluency and proficiency. With respect to ASE, they suggest that articulation rate best reflects the componential structure of functional adequacy and is the best proxy for proficiency if conceptualized as the informativeness and

comprehensibility of learner productions, i.e. functional adequacy. End and mid-clause pause duration do, however, also represent some of the processes that underpin oral proficiency. In line with Skehan et al. (2016) claim that clause fluency is associated with formulation and and discourse-fluency is associated with conceptualization, speed of syntactic processing predicted mid-clause pause duration, and speed of lexical processing predicted end clause pause duration. If the aim is to mimic human raters using current assessment rubrics which comprise a set of analytic scales focusing on different dimensions of oral proficiency, then it might be appropriate to use a combination of mid- and end-clause pause duration and articulation rate as proxies for global measures of proficiency.

That measures of linguistic knowledge and processing only account for a relatively small proportion of variation in measures of utterance fluency is unsurprising given the growing body of research that observes a relationship between L1 and L2 fluency, and suggests that measures of fluency might reflect a speaker's individual speaking style (Bradlow et al., 2017; De Jong et al., 2015; Derwing et al., 2009; Huensch & Tracy-Ventura, 2017; Kahng, 2020; Kim et al., 2013; Peltonen, 2018; Towell & Dewaele, 2005).

CONCLUSION

This study was carried out to address the circularity present in much validation work in the area of speaking assessment and ASE. Measures of utterance fluency, i.e. pace and pausing behaviors, were compared with measures of overall proficiency, operationalized as functional adequacy, as were models of the construct that underpins measures of utterance fluency and models of the construct that underpins oral proficiency.

The main findings of this study were that articulation rate, mean pause duration and repetition frequency are predictors of functional adequacy. Breadth of lexical knowledge is the main predictor of both functional adequacy and articulation rate. Moreover, a similar combination of dimensions of linguistic knowledge and processing appear to predict both functional adequacy and articulation rate.

These results reflect those of previous studies which have observed a strong relationship between articulation rate and oral proficiency (Ginther et al., 2010) and that vocabulary knowledge is a significant predictor of both functional adequacy (De Jong et al., 2012) and measures of utterance fluency (De Jong et al., 2013). Together our results suggest that it is valid to include some measures of utterance fluency, and in particular articulation rate, in ASE models and to reference fluency in the rubrics provided to human raters, more broadly, because these measures seem to reflect the same underlying construct.

It should, however, be noted that together the measures of utterance fluency only accounted for 60% of the variation in functional adequacy scores. As such, measures of utterance fluency may not be considered to provide an adequate representation of oral proficiency and ought to be supplemented with other measures such as measures of goodness of pronunciation, lexical diversity and grammatical accuracy (Higgins et al., 2011; Zechner et al., 2009).

It is also interesting to note that the measures of linguistic knowledge and processing used in this study only accounted for between 6% and 32% variation in measures of utterance fluency. This finding, reflecting other studies that have found that these measures account for between 5% and 50% of variation in utterance

fluency (De Jong et al., 2013), suggests that measures of utterance fluency also represent other factors. The discrepancy between the results of this study and those of De Jong et al. (2013) might be explained by the fact that de Jong et al.'s study also included a measure of speed of articulation, i.e. response duration in a picture naming task. Another possibility is that they reflect a learner's individual speaking style (Bradlow et al., 2017; De Jong et al., 2015; Derwing et al., 2009; Huensch & Tracy-Ventura, 2017; Kahng, 2020; Kim et al., 2013; Peltonen, 2018; Towell & Dewaele, 2005) which has been shown to reflect gender, age, dialect and other demographic and individual factors (Bona, 2014; Dewaele & Furnham, 1999, 2000; Jacewicz et al., 2010). A further possibility is that they reflect task type – this study only employed one task type, namely a descriptive task, compared with De Jong et al. (2013) which employed eight different tasks that varied in terms of complexity, formality and discourse.

In conclusion, while the small sample size (Tabachnick & Fidell, 2012) and focus on learners from a single L1 background of a somewhat restricted proficiency range limits the generalizability of the findings of the current study, the results of the present study provide evidence of the concurrent and construct validity of the use of measures of utterance fluency in ASE scoring models. With respect to concurrent validity, the results confirm that the following measures of utterance fluency reflect oral proficiency operationalized as communicative adequacy: articulation rate, mid-clause pause frequency, and repetition frequency. With respect to construct validity, the results suggest that articulation rate reflects breadth of lexical knowledge, end clause pause duration reflects speed of lexical processing, and mid-clause pause duration reflects speed of syntactic processing. That different measures of utterance fluency reflect different dimensions of oral proficiency also suggests that ASE scoring models ought to include a combination of measures of utterance fluency, including measures of speed and breakdown fluency and that it might be possible to develop diagnostic assessments using ASE technology in the future. It should, however, be noted that the measures of utterance fluency considered in this study accounted for only 60% of the variation in functional adequacy scores. Measures of utterance fluency would therefore need to be supplemented with more direct measures of the quality of learners' oral productions such as measures of goodness of pronunciation, lexical diversity and grammatical accuracy to ensure ASE scoring models adequately represent oral proficiency.

Like most previous research on oral fluency, this study focused on learners' performance in monologic tasks. It is, however, increasingly recognized that monologic tasks underrepresent oral proficiency because they do not test interactional competence, "the ability to contribute to the shared understanding of information by orally responding appropriately to a given situation" (Chukharev-Hudilainen & Ockey, 2021, p. 1). Moreover, examinees believe that dialogic tasks provide a better reflection of their oral proficiency (Brooks & Swain, 2015; Ockey & Li, 2015). Test developers are therefore starting to explore the possibility of automating dialogic tasks for the purpose of language assessment using spoken dialogue systems, i.e. chatbots (see for example Ockey & Chukharev-Hudilainen, 2021). Future research will therefore need to validate the measures of dialogic fluency these automated interactive speaking assessments rely on and, as a prerequisite, operationalize interactional competence (see Peltonen, 2020 for a review of approaches to measuring interactional competence).

Acknowledgments

This study was funded by the British Council and the Department of Education, University of York. The authors would like to acknowledge Dr Ping (Abby) Wang, Dr Maha Alghasab and Dr Xiaoyin Yang for their support with data collection. Acknowledgement should also be given to the following RAs for their support with coding: Dr Khalid Alahmed, Rachel Brown, Dr Stewart Cooper, Dr Sara Ebrahimi, and Emily Severn, and to Dr Sible Andringa for support with statistics funded by a British Academy Skills Acquisition Award.

Disclosure statement

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

Funding

This work was supported by the British Academy [Skills Acquisition Award]; British Council [ELTRA].

ORCID

Zoe L. Handley http://orcid.org/0000-0002-4732-3443
Haiping Wang http://orcid.org/0000-0001-9680-860X

References

Alexander, B., Ashford-Rowe, K., Barajas-Murph, N., Dobbin, G., Knott, J., McCormack, M., & Weber, N. (2019). *Educause Horizon report 2019 higher education edition* (pp. 3–41). EDU19.

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., Ching Lu, C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, K., Gerdjikova, I., Gutierrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., Orozco-Figueroa, A., & Tzeng, A.. . . Herron, D. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review*, *10*(2), 344–380. https://doi.org/10.3758/BF03196494

Bernstein, J., & Cheng, J. (2008). Logic and validation of a fully automatic spoken English test. In M. Holland & F. P. Fisher (Eds.), *The path of speech technologies in computer assisted language learning* (pp. 188–208). Routledge.

Bernstein, J., Cheng, J., & Suzuki, M. (2011). Fluency changes with general progress in l2 proficiency. In *Twelfth Annual Conference of the International Speech Communication Association (ISCA)*. Florence, Italy.

Bernstein, J., De Jong, J. H. A. L., Pisoni, D., & Townshend, B. (2000, August). Two experiments on automatic scoring of spoken language proficiency. In *Proceedings of InSTIL2000: Integrating Speech Technology in Learning* (pp. 57–61). Dundee, UK: International Speech Communication Association.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*(3), 355–377. https://doi.org/10.1177/0265532210364404

Bona, J. (2014). Temporal characteristics of speech: The effect of age and speech style. *The Journal of the Acoustical Society of America*, *136*(2), EL116–EL121. https://doi.org/10.1121/1.4885482

Bradlow, A. R., Kim, M., & Blasingame, M. (2017). Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts L2 speaking rate. *The Journal of the Acoustical Society of America*, *141*(2), 886–899. https://doi.org/10.1121/1.4976044

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing*, *29*(1), 91–108. https://doi.org/10.1177/0265532211411078

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., Lepage, M., & Op de Beeck, H. P. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS One*, *5*(5), e10773. https://doi.org/10.1371/journal.pone.0010773

Brodeur, M. B., Guérard, K., Bouras, M., & Paterson, K. (2014). Bank of standardized stimuli (BOSS) phase II: 930 new normative photos. *PLoS One*, *9*(9), e106953. https://doi.org/10.1371/journal.pone.0106953

Brooks, L., & Swain, M. (2015). Students' voices: The challenge of measuring speaking for academic contexts. In B. Spolsky, O. Inbar-Lourie, & M. Tannenbaum (Eds.), *Challenges for language education and policy: Making space for people* (pp. 65–80). Routledge.

Burchfield, A. L., & Bradlow, A. R. (2014). Syllabic reduction in mandarin and English speech. *The Journal of the Acoustical Society of America*, *135*(6), EL270–EL276. https://doi.org/10.1121/1.4874357

Case. (2008). 101 IELTS speaking part two tasks about people, places, actions, things and times. *Using English.Com*. https://www.usingenglish.com/files/pdf/101-ielts-speaking-part-two-tasks-about-people-places-actions-things-and-times.pdf

Chen, L., Zechner, K., Yoon, S., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., Ma, M., Mundkowsky, R., Lu, C., Leong, C. W., & Gyawali, B. (2018). Automated scoring of nonnative speech using the SpeechRater SM v. 5.0 engine. *ETS Research Report Series*, *2018*(1), 1–31. https://doi.org/10.1002/ets2.12198

Chukharev-Hudilainen, E., & Ockey, G. J. (2021). The development and evaluation of interactional competence elicitor for oral language assessments. *ETS Research Report Series*, *2021*(1), 1–20. https://doi.org/10.1002/ets2.12319

Coupé, C., Oh, Y. M., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, *5*(9), eaaw2594. https://doi.org/10.1126/sciadv.aaw2594

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 113–132. https://doi.org/10.1515/iral-2016-9993

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

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(2), 223–243. https://doi.org/10.1017/S0142716413000210

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5–34. https://doi.org/10.1017/S0272263111000489

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, *34*(5), 893–916. https://doi.org/10.1017/S0142716412000069

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

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31*(4), 533–557. https://doi.org/10.1017/S0272263109990015

Dewaele, J. M., & Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning*, *49*(3), 509–544. https://doi.org/10.1111/0023-8333.00098

Dewaele, J. M., & Furnham, A. (2000). Personality and speech production: A pilot study of second language learners. *Personality and Individual Differences*, *28*(2), 355–365. https://doi.org/10.1016/S0191-8869(99)00106-3

Ebrahimi, S. (2021). *The impact of semantic mapping technique on the organization of bilingual mental lexicon and L2 utterance fluency of Iranian EFL learners* [PhD thesis], University of York.

Ellis, R., & Barkhuizen, G. P. (2005). *Analysing learner language*. Oxford University Press.
Engelhardt, P. E., Veld, S. N., Nigg, J. T., & Ferreira, F. (2012). Are language production problems apparent in adults who no longer meet diagnostic criteria for attention-deficit/hyperactivity disorder? *Cognitive Neuropsychology*, *29*(3), 275–299. https://doi.org/10.1080/02643294.2012.712957
Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly*, *10*(3), 274–291. https://doi.org/10.1080/15434303.2013.769548
Farrell, M. T., & Abrams, L. (2014). Picture–word interference reveals inhibitory effects of syllable frequency on lexical selection. *The Quarterly Journal of Experimental Psychology*, *67*(3), 525–541. https://doi.org/10.1080/17470218.2013.820763
Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, *18*(3), 299–323. https://doi.org/10.1017/S0272263100015047
Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375.
Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379–399. https://doi.org/10.1177/0265532210364407
Handley, Z. L., & Wang, H. (2018). *What is the Impact of Study Abroad on Oral Fluency Development? A Comparison of Study Abroad and Study at Home*. The British Council. https://englishagenda.britishcouncil.org/sites/default/files/attachments/h136_elt_a4_what_is_the_impact_of_study_abroad_on_oral_fluency_development_final_web.pdf
Harsch, C., Ushioda, E., & Ladroue, C. (2017). Investigating the predictive validity of TOEFL iBT® test scores and their use in informing policy in a United Kingdom University setting. *ETS Research Report Series*, *2017*(1), 1–80. https://doi.org/10.1002/ets2.12167
Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, *25*(2), 282–306. https://doi.org/10.1016/j.csl.2010.06.001
Hsieh, C. N., Zechner, K., & Xi, X. (2019). Features measuring fluency and pronunciation. In K. Zechner & K. Evanini(Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 101–122). Routledge.
Huang, C. J. (1988). Wǒpǎo de kuài and Chinese phrase structure. *Language*, *64*(2), 274–311. https://doi.org/10.2307/415435
Huang, C. J. (1989). Pro-drop in Chinese: A generalized control theory. In O. A. Jaeggli & K. J. Safir (Eds.), *The null subject parameter* (pp. 185–214). Springer.
Huensch, A., & Tracy-Ventura, N. (2017). Understanding second language fluency behavior: The effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycholinguistics*, *38*(4), 755–785. https://doi.org/10.1017/S0142716416000424
Hulstijn, J. H., & Schoonen, R. (2004). *What is speaking proficiency (WiSP)*. Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). https://www.narcis.nl/research/RecordID/OND1301810
Hulstijn, J. H., Schoonen, R., De Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the common European framework of reference for languages (CEFR). *Language Testing*, *29*(2), 203–221. https://doi.org/10.1177/0265532211419826
Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49. https://doi.org/10.1093/applin/amm017
Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America*, *128*(2), 839–850. https://doi.org/10.1121/1.3459842

Jaeggli, O., & Safir, K. J. (1989). The null subject parameter and parametric theory. In O. A. Jaeggli & K. J. Safir (Eds.), *The null subject parameter* (pp. 1–44). Springer.

Jared, D., Poh, R. P. Y., & Paivio, A. (2013). L1 and L2 picture naming in mandarin–English bilinguals: A test of bilingual dual coding theory. *Bilingualism: Language and Cognition*, *16*(2), 383–396. https://doi.org/10.1017/S1366728912000685

Jiang, N. (2002). Form–meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, *24*(4), 617–637. https://doi.org/10.1017/S0272263102004047

Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics*, *25*(4), 603–634. https://doi.org/10.1017/S0142716404001298

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, *64*(4), 809–854. https://doi.org/10.1111/lang.12084

Kahng, J. (2020). Explaining second language utterance fluency: Contribution of cognitive fluency and first language utterance fluency. *Applied Psycholinguistics*, *41*(2), 457–480. https://doi.org/10.1017/S0142716420000065

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 18–64). American Council on Education/Praeger.

Kim, M., Ackerman, L., Burchfield, L. A., Dawdy-Hesterberg, L., Luque, J., Mok, K., & Bradlow, A. (2013). Rate variation as a talker-specific/language-general property in bilingual speakers. *The Journal of the Acoustical Society of America*, *133*(5), 3574–3574. https://doi.org/10.1121/1.4806558

Koizumi, R., In'nami, Y., & Jeon, E. H. (2022). L2 speaking and its external correlates: A meta-analysis. In E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency* (pp. 339–367). John Benjamins.

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

Kuiken, F., & Vedder, I. (2018). Assessing functional adequacy of L2 performance in a task-based approach. In N. Taguchi & Y. Kim (Eds.), *Task-based approaches to teaching and assessing pragmatics* (pp. 266–285). John Benjamins.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51. https://doi.org/10.1177/026553229901600103

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.

Levelt, W. J. M. (1999). Producing spoken language: A blueprint of the speaker. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 83–122). Oxford University Press.

Magnan, S. (1988). Grammar and the ACTFL oral proficiency interview: Discussion and data. *The Modern Language Journal*, *72*(3), 266–276. https://doi.org/10.2307/327504

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, *7*(1), 52–76. https://doi.org/10.1177/026553229000700105

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 57–78). EUROSLA.

Nation, P. (2007). The four strands. *International Journal of Innovation in Language Learning and Teaching*, *1*(1), 2–13. https://doi.org/10.2167/illt039.0

Norris, J. M. (2005). Using developmental sequences to estimate ability with English grammar: Preliminary design and investigation of a web-based test. *University of Hawaii Second Language Studies Paper*, *24*(1). http://hdl.handle.net/10125/40678

Ockey, G. J., & Chukharev-Hudilainen, E. (2021). Human versus computer partner in the paired oral discussion test. *Applied Linguistics*, *42*(5), 924–944. https://doi.org/10.1093/applin/amaa067

Ockey, G. J., & Li, Z. (2015). New and not so new methods for assessing oral communication. *Language Value*, *7*(7), 1–21. https://doi.org/10.6035/LanguageV.2015.7.2

Peirce, J. W. (2007). PsychoPy - psychophysics software in Python. *Journal of Neurosciemce Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Peltonen, P. (2018). Exploring connections between first and second language fluency: A mixed methods approach. *The Modern Language Journal*, *102*(4), 676–692. https://doi.org/10.1111/modl.12516

Peltonen, P. (2020). *Individual and Interactional Speech Fluency in L2 English from a Problem-solving Perspective: A Mixed-methods Approach*. Unpublished PhD Thesis. University of Turku.

Porte, G. (Ed.). (2012). *Replication research in applied linguistics*. Cambridge University Press.

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*(3), 355–371. https://doi.org/10.1177/026553229301000308

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined. In B. Laufer & P. Bogaards (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 209–227). John Benjamins.

Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, *37*(6), 828–848. https://doi.org/10.1093/applin/amu069

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *The Canadian Modern Language Review*, *65*(3), 395–412. https://doi.org/10.3138/cmlr.65.3.395

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Segalowitz, N., French, L., & Guay, J. D. (2017). What features best characterize adult second language utterance fluency and what do they reveal about fluency gains in short-term immersion? *Canadian Journal of Applied Linguistics*, *20*(2), 90–116. https://doi.org/10.7202/1050813ar

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. https://doi.org/10.1515/iral-2016-9992

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics*. Pearson.

Tardif, T., Shatz, M., & Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs: A comparison of English, Italian, and mandarin. *Journal of Child Language*, *24*(3), 535–565. https://doi.org/10.1017/S030500099700319X

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal*, *65*(1), 71–79. https://doi.org/10.1093/elt/ccq020

Tavakoli, P., Nakatsuhara, F., & Hunter, A. M. (2017). Scoring validity of the aptis speaking test: Investigating fluency across tasks and levels of proficiency. *ARAGs Research Reports Online*. AR-G/2017/7. ISSN 2057-5203.

Towell, R., & Dewaele, J. M. (2005). The role of psycholinguistic factors in the development of fluency amongst advanced learners of French. In J. M. Dewaele (Ed.), *Focus on French as a foreign language: Multidisciplinary approaches* (pp. 210–239). Multilingual Matters.

van Dalen, R. C., Knill, K. M., & Gales, M. J. (2015). Automatically grading learners' English using a Gaussian process. *Proceedings Sixth Workshop on Speech and Language Technology in Education (SLaTE)*, 7–12. Leipzig, Germany: ISCA. https://www.repository.cam.ac.uk/handle/1810/249186

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29*(3), 325–344. https://doi.org/10.1177/0265532211424478

Wang, X., & Evanini, K. (2019). Features measuring content and discourse coherence. In K. Zechner, & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 138–156). Routledge.

Wang, Y., Gales, M. J. F., Knill, K. M., Kyriakopoulos, K., Malinin, A., van Dalen, R. C., & Rashid, M. (2018). Towards automatic assessment of spontaneous spoken English. *Speech Communication*, *104*, 47–56. https://doi.org/10.1016/j.specom.2018.09.002

Weekes, B. S., Shu, H., Hao, M., Liu, Y., & Tan, L. H. (2007). Predictors of timed picture naming in Chinese. *Behavior Research Methods*, *39*(2), 335–342. https://doi.org/10.3758/BF03193165

West, M. (1953). *A general service list of English words*. Longman.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J.,

Robinson, D., Seidel, D., Spinu, V., & Woo, K.... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement Issues & Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Xi, X. (2007). Evaluating analytic scoring for the TOEFL academic speaking test (TAST) for operational use. *Language Testing*, *24*(2), 251–286. https://doi.org/10.1177/0265532207076365

Xi, X., Schmidgall, J., & Wang, Y. (2016). Chinese users' perceptions of the use of automated scoring for a speaking practice test. In G. Yu & Y. Jin (Eds.), *Assessing Chinese learners of English: Language constructs, consequences and conundrums* (pp. 150–175). Palgrave Macmillan.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215–229.

Yan, X. (2020). Unpacking the relationship between formulaic sequences and speech fluency on elicited imitation tasks: Proficiency level, sentence length, and fluency dimensions. *TESOL Quarterly*, *54*(2), 460–487. https://doi.org/10.1002/tesq.556

Yoon, S. Y., Lu, X., & Zechner, K. (2019). Features measuring vocabulary and grammar. In K. Zechner, and K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 123–137). Routledge.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*(10), 883–895. https://doi.org/10.1016/j.specom.2009.04.009

Appendix A: Vocabulary Processing Task

Instructions

Task 5 Picture Naming Task

In this task, on each trial you will see an image. Your task is to say the name of the object presented in the image. You should try to respond as quickly as possible.

Do not worry if you cannot name all of the objects.

When you are ready to start the experiment, please let the researcher know.

Stimuli

Practice

Dog, banana, camera, bear, butterfly, guitar.

Experimental

Calendar, elephant, envelope, rabbit, telephone, candle, lion, pencil, helicopter, scissors, headphones, lipstick, mushroom, paintbrush, tiger, button

Appendix B: Morphosyntactic Processing Task

Instructions

Task 6 Grammar Task

A On each trial you will be presented a series of words. Your task is to construct as short a sentence as possible using the words presented, and say it aloud. You should try to respond as quickly as possible. Make sure you pay attention to TENSE and NUMBER.

Do not worry if you cannot construct all of the sentences.

For example:

You are presented:		You say:
(everyday)		
Man	to write	*The man writes (everyday)*
(now)		
two + girl	to read	*The two girls are reading (now)*
(yesterday)		
Bank	to close	*The bank closed (yesterday)*
(tomorrow)		
two + waiter	to phone	*The two waiters will phone (tomorrow)*

When you are ready to start the experiment, let the researcher know.

Stimuli

Note: Contractions were also accepted

Also alternative places for the adverb, e.g. “she is now sleeping”.

Table B1. Stimuli in the morphosyntactic processing task.

Stimulus				Correct Answer
P00	He	to study	(everyday)	he studies (everyday)
P01	two + shop	to close	(now)	(the) two shops are closing (now)
P02	They	to shut	(tomorrow)	they will/are going to shut (tomorrow)
P03	Woman	to arrive	(yesterday)	the/a woman arrived (yesterday)
P04	two + boy	to cook	(yesterday)	(the) two boys cooked (yesterday)
P05	She	to sing	(tomorrow)	she will/is going to sing (tomorrow)
E00	He	to write	(everyday)	he writes (everyday)
E01	two + woman	to kneel	(now)	(the) two women are kneeling (now)
E02	They	to ring	(tomorrow)	they will/are going to ring (tomorrow)
E03	Glass	to crack	(yesterday)	the/a glass cracked (yesterday)
E04	Artist	to paint	(everyday)	the/an artist paints (everyday)
E05	She	to sleep	(now)	she is sleeping (now)
E06	two + team	to win	(tomorrow)	(the) two teams will/are going to win (tomorrow)
E07	They	to clean	(yesterday)	they cleaned (yesterday)
E08	They	to fight	(everyday)	they fight (everyday)
E09	Flower	to grow	(now)	the/a flower is growing (now)
E10	He	to wash	(tomorrow)	he will/is going to wash (tomorrow)
E11	two + boat	to sink	(yesterday)	(the) two boats sank (yesterday)
E12	two + girl	to run	(everyday)	(the) two girls run (everyday)
E13	They	to play	(now)	they are playing (now)
E14	Plane	to crash	(tomorrow)	the/a plane will is going to crash (tomorrow)
E15	She	to come	(yesterday)	she came (yesterday)

Appendix C: Syntactic Processing Task

Instructions

Task 7 Grammar Task B

On each trial you will be presented a series of words. Your task is to construct as short a sentence as possible using the words presented, and say it aloud. You should try to respond as quickly as possible. Make sure you pay attention to whether you need to form a POSITIVE or a NEGATIVE QUESTION or STATEMENT.

Do not worry if you cannot construct all of the sentences.

For example:

You are presented:			You say:
	+		
the milk	the cat	is drinking	*The cat is drinking the milk*
	-		
the milk	the cat	is drinking	*The cat is not drinking the milk*
	?		
the milk	the cat	is drinking	*Is the cat drinking the milk?*

When you are ready to start the experiment, let the researcher know.

Stimuli

The following were also accepted: contractions, lexical substitutions and mispronunciations, alternative negatives, and alternative negatives. Errors in conjugation (morpho-syntax) were ignored given the focus of this test on syntax.

Table C1. Stimuli in the syntactic processing task.

Stimulus					Correct Answer
P00	Statement	the man	the door	is closing	the man is closing the door
P01	Negative	the paper	the men	are tearing	the men are not tearing the paper
P02	Question	the horse	the gate	Kicked	did the horse kick the gate
P03	Statement	the cakes	the women	Baked	the women baked the cakes
P04	Question	television	the man	is watching	is the man watching the television
P05	Negative	the books	the women	Read	the women did not read the books
E00	Statement	the dogs	the bones	are hiding	the dogs are hiding the bones
E01	Negative	the girls	the bikes	are riding	the girls are not riding the bikes
E02	Question	the women	the buckets	are carrying	are the women carrying the buckets
E03	Statement	the cat	the food	Found	the cat found the food
E04	Negative	the men	the drums	Played	the men did not play the drums
E05	Question	the goats	the flowers	Ate	did the goats eat the flowers
E06	Statement	the houses	the children	are passing	the children are passing the houses
E07	Negative	the pipes	the men	are smoking	the men are not smoking the pipes
E08	Question	the combs	the girls	are using	are the girls using the combs
E09	Statement	the jackets	the men	Wore	the men wore the jackets
E10	Negative	the babies	the girls	Dressed	the girls did not dress the babies
E11	Question	the chains	the bears	Broke	did the bears break the chains
E12	Statement	the cook	the boat	is describing	the cook is describing the boat
E13	Negative	the boy	the egg	is painting	the boy is not painting the egg
E14	Question	the woman	the car	is renting	is the woman renting the car
E15	Statement	the doctor	the question	Asked	the doctor asked the question
E16	Negative	the actor	the hat	Bought	the actor did not buy the hat
E17	Question	the sailor	the star	Saw	did the sailor see the star
E18	Statement	the plane	the man	Flew	the man flew the plane
E19	Negative	the road	the woman	is clearing	the woman is not clearing the road
E20	Question	the ball	the cat	is watching	is the cat watching the ball
E21	Statement	the bottle	the man	Shook	the man shook the bottle
E22	Negative	the box	the woman	Opened	the woman did not open the box
E23	Question	the toy	the baby	Wanted	did the baby want the toy