Aspects of fluency across assessed levels of speaking proficiency

Article

Accepted Version

Aspects of fluency across assessed levels of speaking proficiency

Authors: Parvaneh Tavakoli, Fumiyo Nakatsuhara, Ann-Marie Hunter

Abstract

Recent research in second language acquisition suggests that a number of speed, breakdown, repair and composite measures reliably assess fluency and predict proficiency. However, there is little research evidence to indicate which measures best characterize fluency at each assessed level of proficiency, and which can consistently distinguish one level from the next. This study investigated fluency in 32 speakers' performing four tasks of the British Council's Aptis Speaking test, which were awarded four different levels of proficiency (CEFR A2-C1). Using PRAAT, the performances were analysed for various aspects of utterance fluency across different levels of proficiency. The results suggest that speed and composite measures consistently distinguish fluency from the lowest to upper-intermediate levels (A2-B2), and many breakdown measures differentiate between the lowest level (A2) and the rest of the proficiency groups, with a few differentiating between lower (A2, B1) and higher levels (B2, C1). The varied use of repair measures at different levels suggest that a more complex process is at play. The findings imply that a detailed micro-analysis of fluency offers a more reliable understanding of the construct and its relationship with assessment of proficiency.

Keywords: fluency; measuring fluency; assessing fluency; proficiency; rating scale validation

INTRODUCTION

The interest in researching second language (L2) oral fluency has increased over the past decades due to the significant role it plays in reflecting the development of communicative

ability and assessing learner proficiency. From a developmental perspective, when L2 learners expand their L2 repertoire and progress to a higher proficiency level, their language use becomes more automatic and they typically produce output of higher fluency, accuracy and complexity (Housen, Kuiken, & Vedder, 2012; Skehan, 2009, 2015). The concept of fluency, i.e., producing language at an adequate speed with more ease and less hesitation, is therefore related to proficiency level where a positive correlation is usually assumed between the two[i]. In L2 assessment 'fluency' has long been recognized a key construct that reflects L2 proficiency (see Fulcher 2003, pp. 9-10), and as such, it is included in rating scales of speaking exams (e.g., Cambridge General English exams, IELTS, TOEFL iBT) and language benchmarks for L2 communicative ability (e.g., CEFR; Council of Europe, 2001).

Researchers commonly agree that fluency is a complex and multifaceted construct, often difficult to define and measure (Kormos, 2006; Lennon, 1990; Segalowitz, 2010). In recent years, however, attempts have been made to unpack the concept of fluency, and to identify ways of measuring it reliably. Segalowitz's (2010) model of fluency and Skehan's (2003) framework for measuring it are two examples of successful attempts that have expanded our conceptual understanding of fluency, providing the discipline with more valid and reliable indices of fluency. Following from Skehan (2003), researchers in this area have reported that fluency can be consistently measured using indices related to three key aspects of fluency: speed, breakdown, and repair, i.e. how fast a speaker talks; how much pausing disrupts the flow of speech, and how much repair is used to correct, reformulate and restore L2 utterances (Kahng, 2014; Kormos, 2006; Tavakoli & Skehan, 2005). More recently, other researchers (Huensch & Tracy-Ventura, 2016; Hunter, 2017; Skehan, 2015; Bosker et al., 2013) have argued that a distinction can be made between (a) composite measures of fluency, i.e. measures that combine two or more of these aspects (e.g., speed *and* breakdown fluency in 'Mean Length of Run', and speed, breakdown *and* repair in the measure 'Pruned Speech Rate'), and (b) pure

measures, i.e., measures that examine only one aspect (Skehan, 2014). The rationale supporting the distinction is that the two sets of measures may have different strengths and weaknesses, and therefore may be useful for different purposes. For example, while composite measures correspond more strongly with human judgement of fluency (e.g., Kormos & Denes, 2004), pure measures tell us more about the underlying processes of speech formulation and production (e.g., Huensch & Tracy-Ventura, 2017) and can therefore provide a more nuanced picture of fluency that is more appropriate in language teaching and assessment (Tavakoli & Hunter, 2018). From a psycholinguistic perspective, an in-depth and detailed analysis of L2 fluency is also perceived as crucial, since the emerging evidence (de Jong, et al., 2015; Derwing et al., 2009; Huensch & Tracy-Ventura, 2017; Peltonen, 2018) suggests that L1 and L2 fluency behaviours are, at least to some extent, related, and that certain aspects of L2 fluency might be a function of L1 personal styles. In sum, then, given this complex picture, it can be argued that conceptualizing and measuring fluency at a fine-grained level can not only reveal more about the connection between L2 speech and the underlying speech production processes (Huensch & Tracy-Ventura, 2017; Hunter, 2017; Tavakoli & Hunter, 2018), but it will enhance a more reliable understanding of what characterizes fluency at different levels of proficiency, making speaking test rating scales more useful and meaningful for users and examiners (Nakatsuhara, 2014; Tavakoli, et al., 2017).

In language assessment, there is robust evidence to indicate that fluency relates to communicative adequacy (de Jong et al., 2015; Revesz, Ekiert, & Torgersen, 2016), affects raters' perceptions of L2 ability (Prefontaine, Kormos, & Johnson, 2016) and predicts proficiency (Iwashita et al., 2008; Revesz et al., 2016). Recent research in this area (Nakatsuhara, 2014; Tavakoli et al., 2017) has also indicated that, despite its significance in the assessment of proficiency and its impact on raters' judgements and ratings, fluency is a relatively under-researched area. While it is commonly believed that fluency increases with

proficiency, perhaps because they both develop as a result of language use and practice, there is little evidence to demonstrate this relationship is linear for all different features of fluency under four main aspects of fluency, i.e., speed, breakdown, repair, and composite. Nevertheless, it appears that fluency descriptors in many operational fluency scales seem to assume that all features develop linearly. From a measurement perspective, fluency representations in language examination descriptors are rather limited and inconsistent, failing either to consider all key aspects of fluency or to pinpoint the individual contribution of these different aspects to the ratings of fluency. These are some of the gaps the current study aims to help fill by investigating fluency in a detailed micro-analytic framework, across assessed levels of proficiency in the Aptis Speaking test.

THEORETICAL BACKGROUND

Understanding fluency

Segalowitz (2010) offers a triadic framework for understanding and defining fluency. In this framework, L2 fluency is presented as three different but inter-related 'domains' of *cognitive*, *utterance*, and *perceived* fluency. *Cognitive fluency*, according to Segalowitz (2010, p. 202), is "the efficiency of the operation of the cognitive mechanisms underlying performance" and "the ability to efficiently mobilize and integrate the underlying cognitive processes responsible for producing utterances" (Segalowitz, 2010, p. 48). *Utterance fluency* relates to the measurable aspects of fluency in uttered speech such as speed, pausing and hesitation, and *perceived fluency* represents the inferences listeners make about speakers' cognitive fluency based on their perceptions of the speech that they hear. Of the three components of the framework, utterance fluency has been widely investigated as it is particularly amenable to quantitative methods and allows for a degree of standardisation and comparison across studies.

Researching L2 fluency is particularly important for the development of a model of L2 speech production. Many SLA researchers (e.g., Kahng, 2014; Kormos, 2006; Skehan, 2009), have relied on Levelt's (1989, 1999; see also Field, 2011) L1 model hypothesizing that speech is processed and produced through four different stages of *conceptualization, formulation* (including grammatical encoding, morpho-phonological encoding, phonetic encoding), *articulation,* and *self-monitoring*. During conceptualization, according to Levelt, the speaker generates a pre-verbal message which is then transferred to the formulation stage, where the original message is shaped and formulated through grammatical, morpho-phonological and phonetic encoding of lemmas accessed in the lexicon. This linguistic message is then transferred to the articulation stage, where the linguistic plan turns into actual speech. An additional stage is self-monitoring, where speakers' attention is focussed on the message both during formulation and shortly after it is uttered to check for accuracy, clarity and appropriacy. Kormos (2006) argues that a key distinction between L1 and L2 speech production processes is that L2 speakers must access their declarative knowledge of syntactic and phonological rules, and that their language processing may be less automatic.

Measuring fluency

Over the past decades, a number of developments have come about in the operationalization and measurement of fluency. It is now common to see a distinction made between "pure" and "composite" measures. These pure measures relate to each of the three aspects of fluency: speed, breakdown, and repair. For example, a researcher can isolate the speed with which a person speaks by calculating the speaker's *Articulation Rate.* Similarly, the researcher can investigate breakdown and repair fluency in isolation by counting frequency of pauses or frequency of reformulations, respectively. Composite measures, on the other hand, consider two or more of the aspects in combination. For example, if we find the average number of

syllables produced in runs of speech between silences (*Mean Length of Run*), this could tell us something about the extent to which a speaker is silent (breakdown fluency) but also how many syllables they are able to produce during that run (speed fluency). In other words, if a speaker has a high mean length of run it could be because they pause very infrequently and/or because they produce a high volume of syllables between pauses. Similarly, *Speech Rate* will be high for a speaker who fills most of the sample time with speech and avoids hefty pauses (breakdown fluency) but will also be high for someone who produces a large number of syllables (i.e. speaks quickly). Because the process of pruning speech often involves removing syllables involved in repairing an utterance (i.e. a reformulation or self-correction), a measure which uses pruned speech (such as, *Pruned Speech Rate*) could be said to combine all three aspects of fluency.

In terms of 'pure' measures, firstly, the *speed* with which a person speaks is sometimes considered reflective of the articulatory process only (de Jong & Perfetti, 2011; Huensch & Tracy-Ventura, 2017), but it has also been argued that it gives an indication of the extent to which the speaker "buys time" while speaking by lengthening sounds (Hilton, 2014; Hunter, 2017). Put simply, the speed of L2 speech is believed to be related to the speaker's "buying time" during speech to either plan the upcoming utterance (conceptualization) or carry out lexical, morphosyntactic and phonological encoding (formulation/articulation). Another important development in speed fluency measurement is the use of sophisticated computer software, such as PRAAT, which enables researchers to calculate syllable counts and duration of runs of speech quickly and more reliably.

While there is arguably only one way to measure speed (articulation rate/syllable length)[ii], *breakdown* fluency can be examined in various ways which take into consideration the *amount*, *location* and *character* of the pausing. *Amount* of pause can be gauged using a measure such as phonation time ratio which calculates the proportion of time spent speaking and silent. Breakdown can also be looked at in more detail in terms of the *length* or *frequency*

of pauses. Bosker et al., (2013) show that pause frequency is likely to be a more important indicator of L2 breakdown fluency than pause length. Similarly, Prefontaine (2013) contends that listeners will tolerate very long pauses, provided that they come at an acceptable point in the utterance (i.e., at a clause boundary). The evidence about the importance of the *location* of pauses, i.e., whether pauses occur in mid-clause or end-clause positions, initially came from studies such as Tavakoli (2011) who argued that the distinction between L1 and L2 speakers' pausing behaviour is not in *how much* they paused, but in *where* they paused. This has been further explored in a number of studies (de Jong, 2016a; Kahng, 2014; Skehan & Foster, 2008). The main argument here is that while L1 speakers pause to process language at the conceptualization phase to work on the pre-verbal message, L2 speakers may rely on mid-clause pauses to formulate the message in terms of lexical or morphosyntactic features (Huensch & Tracy-Ventura, 2017; Skehan & Shum, 2017). Likewise, Skehan, Foster and Shum (2016) argue that mid-clause pauses are related to Levelt's 'formulation' stage, while end-clause pauses are more likely to be related to the 'conceptualization' stage. Another important development in relation to measuring breakdown fluency concerns the *character* of pauses, i.e., whether pauses are filled and unfilled (silent). Clark and Fox Tree (2002) and Schmid and Beers Fägersten (2010) contend that while both types indicate language processing demands, filled pauses may also highlight emphasis, discourse organisation and communication strategies, and therefore facilitate communication (Dewaele, 1996).

From the discussion above, it can be seen that a range of speed and breakdown measures are frequently examined in fluency research. Despite the wide range, there has been little agreement about which measures can best characterize speakers' fluency, or whether certain measures are more relevant to describing fluency in different L2 tasks and contexts or at different stages of L2 development. In a recent study, Segalowitz, French, and Guay (2017) examined a corpus of 100 speakers of French as a second language in an immersion programme

to narrow down a large number of features commonly used in fluency research to a smaller core set that can help operationally define fluency. The results suggested that four measures formed the core of fluency features reflecting "a common underlying fluency construct" (Segalowitz, et al., 2017. p. 100). These were: Mean length of run (number of syllables between silent pauses), mean length of phonation run (seconds of phonation between silent pauses), syllable duration (which can also be expressed as 'inverse articulation rate'), and mean length of silent pauses. While these results are central to conceptualizing and measuring fluency as a construct, they do not reveal much about the relationship between fluency and assessed levels of proficiency as the study did not investigate the participants' proficiency level. Given the importance of pause *location* in understanding fluency, it is surprizing that the study did not examine length or frequency of mid-clause pauses in the data set.

Repair seems to be the most controversial aspect of the fluency triad, and it is often reported to have a complex relationship with perceived fluency (e.g., Bosker et al, 2013). Repair measures may be linked to processes of self-monitoring (Huensch & Tracy-Ventura, 2017; Hunter, 2017) in that a speaker who repairs an utterance is often also monitoring his/her output and making amendments. The extent to which an L2 speaker repairs his or her speech is typically calculated by counting the number of reformulations, false starts, self-corrections, repetitions, replacements or hesitations per 60 seconds. While this is a widely-used approach to measuring repair fluency, a number of limitations can be identified. First, some of these measures overlap with one another or with other aspects of performance, e.g., complexity. False starts, for example, often lead to reformulations (Hunter, 2017), and, therefore, number of false starts and reformulations could be internally dependent. Similarly, hesitations often proceed or co-occur with other repair measures such as repetitions and replacements, and therefore calculating one of these measures would inevitably overlap with others. As such, calculating each measure separately may not provide an independent representation of the

repair phenomenon. Second, it has been argued that verbatim repetitions may not reflect 'repair' behaviour, but, rather, *breakdown*, as repetition may be used by a speaker to stall for time (Dornyei & Kormos, 1998; Witton-Davies, 2014). A final limitation is that, as some emerging research evidence suggests, use of repair measures may be particularly strongly linked with personal speaking styles (e.g., de Jong et al., 2015) although links have also been found between other aspects of fluency and L1 speaking style (e.g., Bradlow et al., 2017; de Jong et al., 2015; Derwing et al., 2009; Huensch & Tracy-Ventura, 2017). While the scope of the current study would not allow us to operationalize fluency-related personal styles or to investigate speakers' rationale in using certain repair measures (e.g., through stimulated recall), we aim to look in detail at the construct of repair fluency by isolating different repair types and choosing measures that are independent of one another.

Developing and validating fluency rating scales

Fluency is one of the most common criteria featured in both holistic and analytic rating scales in many standardized tests of speaking, e.g., Cambridge General English tests, IELTS, TOEFL iBT (de Jong, 2018). Aspects of fluency featured in these rating scales include: length of speech, hesitation, repetition, self-correction, flow of speech, pauses, speed of speech, rhythm, false starts, evenness of speech. Some tests also refer to the underlying cause of the hesitation, for example whether it is content-related or language-related. Depending on the construct of a test, fluency features can be combined with other linguistic features to form one analytic scale (e.g., 'Fluency and Coherence' in IELTS, 'Delivery' including both fluency and pronunciation in TOEFL iBT). While there seems to be a general consensus about the importance of fluency as a key component of L2 proficiency, the fluency descriptors in speaking tests tend to receive limited empirical validation and often appear to have been developed intuitively (de Jong,

2016b; de Jong, 2018; Tavakoli et al., 2017). The existing evidence in publicly available rating descriptors suggests that fluency is usually assessed in a rather limited or ambiguous way, with only few of its fundamental aspects presented in the rating scales (de Jong, 2016b; Tavakoli et al., 2017).

The other concern about assessment of speaking is that human raters' scores are at least partly subjective since they reflect what the raters consider to be 'fluent' or 'dysfluent.' Raters, although often professionally trained and well-experienced, typically work with ambiguous and intuitively-developed rating scales to assess speech samples. The ambiguity of rating scales and the subjective nature of rating processes are two sources of validity and reliability threat that language tests have historically been concerned about. The last few decades have, however, seen several studies attempting to analyse test-takers' spoken performance for two main research objectives: a) to develop and validate speaking test rating scales (e.g., Brown, 2006a; Fulcher, 1996; Iwashita et al., 2008; Nakatsuhara, 2014), and b) to investigate raters' perceptions of proficiency when rating spoken performances (e.g., Brown, 2006b; Brown, Iwashita & McNamara, 2005; Ducasse & Brown, 2009; May, 2011; Pollitt & Murray, 1996). The former type of research conducting a micro-analysis of test-taker performances has consistently suggested that test-takers' progress in different aspects of language (including fluency features) is not linear, indicating that specific aspects of performance are more relevant to differentiate particular levels of proficiency than others.

Iwashita et al. (2008), as a part of a large-scale validation study of TOEFL iBT Speaking test (Brown et al., 2005), examined characteristics of spoken language at five different levels of proficiency from 200 test-takers. The researchers ran ANOVAs with several measures of fluency including the number of filled and unfilled pauses, total pausing time, number of repairs, speech rate, and mean length of run. The results indicated significant differences across proficiency levels for speech rate, unfilled pauses, and total pausing time,

with medium or small effect sizes. Ginther, Dimova, and Yang (2010) investigated selected temporal fluency measures in relation to holistic ratings in the Oral English Proficiency Test (OEPT). Their analysis of 150 test-takers' speech samples indicated strong correlations between the speaking scores and speech rate, speech time ratio, mean length of run, and moderate correlations between the scores and the number and length of silent pauses. However, it was also found that fluency features alone did not appear to distinguish adjacent levels of the OEPT scale.

More recently, Nakatsuhara (2014) used a similar design to investigate fluency characteristics when developing new rating scales for the TEAP (Test of English for Academic Purposes) Speaking test. A range of fluency measures including the number of silent pauses, total pause time, speech rate and articulation rate were compared across three proficiency groups. Although the small sample size of the study (N=23) did not allow for inferential statistics, the means of the three proficiency groups on all fluency measures progressed as the test designers intended, offering *a priori* evidence to the development of the rating scales. The last study relevant to our work is Baker-Smemoe et al. (2014) who used excerpts from 126 ACTFL Oral Proficiency Interviews (OPIs) to investigate whether utterance fluency could predict overall fluency. The fluency measures they targeted were: number and length of pauses, number of hesitations and false starts and mean length of run, speech rate and pruned speech rate. The results suggested that number of pauses, mean length of run and speech rate were able to distinguish between the higher levels of proficiency. It is necessary to highlight that in these studies, assessment of proficiency includes a dimension of fluency, either as an independent fluency construct or in combination with other aspects of performance, e.g., pronunciation or delivery. Given the validation nature of such studies, this kind of design allows researchers to develop an insight into what are considered as characteristics of successful performance at each

proficiency level. Table 1 summarizes the measures and limitations of the four studies reviewed above.

As shown on Table 1, there are several limitations in the studies conducted in this area including the limited choice of fluency measures or proficiency levels under investigation. In particular, these studies have failed to provide a careful analysis of breakdown fluency in terms of pause character (whether silent or filled) and pause location (whether mid or end-clause). Given the potential contribution of different fluency features to helping develop an in-depth understanding of fluency representation, we aim to examine a wide range of fluency features in this study.

TABLE 1: Recent Studies that Examined Fluency Features to Develop/Validate Speaking Tests

	Composite	**Speed**	**Breakdown**	**Repair**	**Limitations**
Iwashita et al (2008)	-Pruned speech rate -Unpruned speech rate -Mean length of run	N/A	-Filled pauses -Unfilled pauses -Total pause time	-Repetition, false starts and reformulations per 60 secs	-No pure speed measures -1 second pause threshold -No mid/end clause pause distinction -No post-hoc comparisons between levels
Ginther et al (2010)	-Speech rate -Mean length of utterance	-Articulation rate	-Silent pause time -Number silent pauses -Mean silent pause -Silent pause ratio -Filled pause time -Number filled pauses -Mean filled pause -Filled pause ratio	N/A	-No repair measures -No mid/end clause pause distinction
Nakatsuhara (2014)	-Pruned speech rate	-Articulation rate	-Number of unfilled pauses per 50 words -Ratio of pause time to speech time	-Ratio of repairs to AS-units	-No mid/end clause pause distinction -No inferential statistics
Baker-Smemoe et al (2014)	-Speech rate -Pruned speech rate -Mean length of run	N/A	-Number and length of pauses	-Number of hesitations -Number of false starts	-No pure speed measures -No mid/end clause pause distinction

RESEARCH AIMS AND QUESTIONS

The prime aim of the current study is to explore the extent to which different aspects of utterance fluency characterize assessed levels of proficiency (A2 to C1 CEFR) in the British Council's *Aptis* Speaking test. The study is specifically interested in investigating whether particular features of fluency are related to assessed levels of proficiency in this test, which provides useful information to the test provider to validate or modify their rating scale descriptors. More importantly, the findings will offer a broader implication about the relationship between different aspects of fluency and assessed levels of proficiency. Assessed levels of speaking proficiency is therefore the independent variable, and analytic measures of utterance fluency are the dependent variables of the study. Proficiency in the current study is represented by the assessment of the spoken samples of the candidates' performance by the British Council's trained raters (see Research Design below for further details). The research question guiding our study is:

Research question (RQ): To what extent can the following aspects of fluency differentiate between different levels of proficiency (A2, B1, B2, and C1 in the CEFR)?

RQ1: Speed fluency

RQ2: Composite fluency

RQ3: Breakdown fluency

RQ4: Repair fluency

METHODOLOGY

Test Tasks

The Aptis Speaking test is a computer-based speaking test consisting of four tasks, each of which targets different CEFR levels from A2 to C1 (see Table 2). The entire test takes about 12 minutes to complete, but the total response time per test is 8 minutes as illustrated in Table 2. Each test-taker's performance on individual test parts is separately examined by different trained raters (e.g., Test-taker A's Part 1 performance is assessed by Rater X, Test-taker A's Part 2 performance is assessed by Rater Y). To enable this rating procedure, individual task performances are recorded in separate audio files, which are then randomized before being distributed to raters. The test-takers are also asked to give their consent for their recordings to be used for research as well as rating purposes. Furthermore, the rating system is innovative, as three different holistic rating scales are used for the target proficiency level of each part (i.e., one scale for Part 1, one for Parts 2 and 3, one for Part 4); this allows raters to provide more accurate ratings for performance at each part. The total scores are calculated by amalgamating the four part-scores obtained by four different raters, which are then converted to appropriate CEFR levels (see O'Sullivan & Dunlea, 2015 for more information).

Table 2: Structure of the Aptis Speaking Test

Part	Description of the tasks	Target level	Rating scale	Response Time
1	Respond to 3 questions on personal topics	A1/A2	A	30 secs x 3
2	Respond to 3 questions, including describing a photo and answering a concrete, familiar topic related to the photo.	B1	B	45 secs x 3
3	Respond to 3 questions related to 2 contrasting pictures.	B1		45 secs x 3
4	Provide a long turn, integrating responses to a set of 3 questions.	B2	C	2 mins (+ 1 min prep.)

While face-to-face speaking tests are often capable of tapping into a wider speaking construct by affording various formats within a test, such as an interview, monologue, role play and paired/group discussion, computer-based speaking tests tend to assess a narrower construct

due to limitations posed by the delivery mode (Nakatsuhara, Berry, Inoue & Galaczi, 2017). The Aptis Speaking test is not an exception, since the four tasks are designed within the limited range of the speaking construct that can be assessed in semi-direct format. That is, the four Aptis tasks are not distinctively different in their structure and cognitive demands, and it was therefore considered that the four tasks would not lead to differences in test-takers' fluency of performance. To examine whether fluency of performance in these different parts of the test was affected by task type, we ran a series of ANOVAs. The non-significant results for all fluency measures across the tasks confirmed that task type in Aptis had no significant effects on fluency of the test-takers' performances (see Tavakoli, et al., 2017 for further details).

Research design

The study had a between-participant design with level of proficiency as a between-participant variable (4 levels of A2 to C1). Thirty-two test-takers, 8 at each level (A2, B1, B2 and C1), were chosen from a large set of operational test recordings that were provided by the examination board. Two experienced Aptis trainers selected speakers whose performance are as typical as possible of the Aptis test-taker population at each level; this helped researchers avoid unusual profiles. This selection process was in addition to the part and total scores awarded in operational test, i.e., allocating both a holistic score and analytical points (e.g., fluency, vocabulary, grammar) to each performance to be weighted equally.

Special care was also taken for a balanced selection of other test-taker characteristics across the four levels as much as possible. Each level included four males and four females. Although Aptis does not gather test-takers' L1 information, it was assumed that the data contained approximately 15 different L1 backgrounds, judging from the 18 countries where the 32 tests were taken. The speculated L1s include Arabic, Bengali, Georgian, German, Japanese,

Spanish, Ukrainian, Uzbek and other seven languages. None of the L1s dominated any of the four levels.

Analytic fluency measures

We are aiming for a detailed analysis framework that can help close gaps or address inconsistencies in the ways the different characteristics of fluency are described at different levels of proficiency. In our analysis, we draw on the recommendations in SLA literature (e.g., de Jong et al., 2015; Kahng, 2014; Skehan, 2014) to select a range of analytic measures that are reported to represent fluency reliably and consistently. The measures that we chose for the initial multivariate analysis are:

Speed measure. Articulation rate: total number of syllables divided by total amount of phonation time (excluding pauses) multiplied by 60[iii]

Composite measure. Speech rate (pruned): total number of syllables divided by total performance time (including pauses) multiplied by 60.

Breakdown measures. Mean length of mid-clause and end-clause pauses

Repair measure. Frequency of all repairs (per 60 seconds speaking time)

It is important to note that for the purpose of the current study we distinguish between a pure measure of speed, i.e., articulation rate and composite measures that combine speed and pausing, i.e., speech rate. We follow Foster, Tonkyn, and Wigglesworth (2000) to divide the speech samples to AS-units and dependent and independent clauses.

Data analysis procedures

All the speech data were transcribed, and a detailed micro-analysis of fluency measures was used to examine the participants' performances. To achieve accurate measurement of fluency, the PRAAT software (Boersma & Weenink, 2013) was used. The speech data were examined and annotated manually by one of the authors. Annotation involved inspecting the spectrogram produced in PRAAT and simultaneously listening to the extracts of speech in order to identify and tag speech phenomena (pauses; repetitions, etc.) on a corresponding grid. Following recent research (e.g., de Jong et al., 2012; de Jong & Bosker, 2013) a minimum pause threshold of 0.25 seconds was set. A computer script was then developed to read the annotated extracts of speech and extract the relevant information needed to calculate the above measures. All the data was annotated for a second time by the same researcher. The intra-rater reliability coefficient of above .90 was achieved across all fluency measures.

RESULTS

A multivariate analysis was run to investigate the effects of Level of proficiency on the participants' performances in terms of the five fluency measures discussed above. Effect sizes were calculated to examine the power of significant results. The results suggested that there were no multivariate outliers in the dependent variables. Levene's Test of Equality of Error Variances showed that the assumption of equality of variance was not violated. The multivariate test showed a statistically significant difference for Proficiency Level (Wilks' Lambda = .225; $F\ (3, 32) = 14.80$, $p < .001$; $\eta^2 = .392$).

A test of between-participant comparisons (Proficiency Level) showed five significant differences for Proficiency level: articulation rate ($F\ (3, 32) = 34.18$, $p < .001$; $\eta^2 = .469$), speech rate ($F\ (3, 32) = 67.97$, $p < .001$; $\eta^2 = .637$), mean length of mid-clause pauses ($F\ (3, 32) = 43.59$, $p < .001$; $\eta^2 = .530$), mean length of end-clause pauses ($F\ (3, 32) = 43.06$, $p <$

.001; η^2 = .527), and total repair (F (3, 32) = 4.78, p < .004; η^2 = .110). These results suggested that further analyses, i.e., ANOVAs[iv], could be used to identify where the statistical differences were in the various measures across proficiency levels.

ANOVAs: Comparisons across proficiency levels

A number of ANOVAs were run to explore the effects of proficiency level on different aspects of fluency. When a significant result was observed, Tukey post-hoc comparison was used to identify where the significant results were located. In this part of the analysis, we extended our focus to include a comprehensive micro-analysis of fluency measures that can help develop a more in-depth understanding of the construct. Given the exploratory nature of this study, it was deemed necessary to run a detailed analysis to see in what ways different aspects of fluency interact with proficiency level. This is an important contribution the current study is aiming to make. To decrease the chance of committing a Type I error, a corrected alpha level of .012 (.05 divided by four) was set. Descriptive statistics is provided for all measures. When significant results are achieved, we provide a figure to demonstrate the comparisons. Since the sample size of the study is relatively small (N=32), the inferential statistics should be interpreted with caution. The measures of fluency analysis were:

TABLE 3: All Measures of Fluency Analysis

Aspects of fluency	**Fluency features**
Speed	Articulation rate
Composite	Speech rate Mean length of run (pruned): the mean number of syllables between two pauses
Breakdown	Phonation time ratio: percentage of performance time spent speaking[v] Mean length of all pauses (filled and silent) Mean length of all silent pauses

	Mean length of silent pauses at mid-clause and end-clause positions, respectively
	Mean length of filled pauses at mid-clause and end-clause positions, respectively
	Frequency of all pauses (filled and silent)
	Frequency of all silent pauses
	Frequency of all filled pauses
	Frequency of silent pauses at mid-clause and end-clause positions, respectively
	Frequency of filled pauses at mid-clause and end-clause positions, respectively
Repair	Frequency of all repairs (per 60 seconds)
	Frequency of false starts and reformulations (per 60 seconds)
	Frequency of partial or complete repetitions (per 60 seconds)
	Frequency of self-corrections (per 60 seconds)

Speed and composite measures

Table 4 shows the means and standard deviations for measures of articulation rate, speech rate and mean length of run across the four levels of proficiency.

TABLE 4: Descriptive Statistics for Speed and Composite Measures Across Proficiency Levels

	Articulation Rate (speed)		**Speech Rate (composite)**		**Mean Length of Run (composite)**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
A2	158.05	23.22	73.24	18.35	3.21	.73
B1	188.04	30.00	135.21	31.57	5.84	1.88
B2	224.25	33.53	172.06	25.81	8.54	1.91
C1	234.90	36.27	172.18	35.51	7.75	1.80

N = 32

Three significant results were observed for speed and composite fluency across different levels of proficiency: one for articulation rate ($F = 34.19$, $p < .001$, $\eta^2 = .467$), one for speech rate ($F = 67.97$, $p < .001$, $\eta^2 = .628$), and one for mean length of run ($F = 52.56$, p

< .001, η^2 = .571). The post-hoc analysis showed that for all these significant results, A2, B1 levels were different from each other and from B2 and C1. However, B2 and C1 levels were not statistically different. Figures 1, 2 and 3 show the results.

FIGURE 1

Articulation Rate

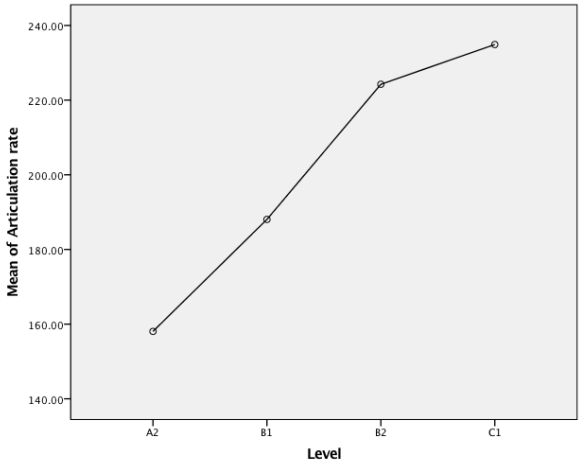


FIGURE 2

Speech Rate

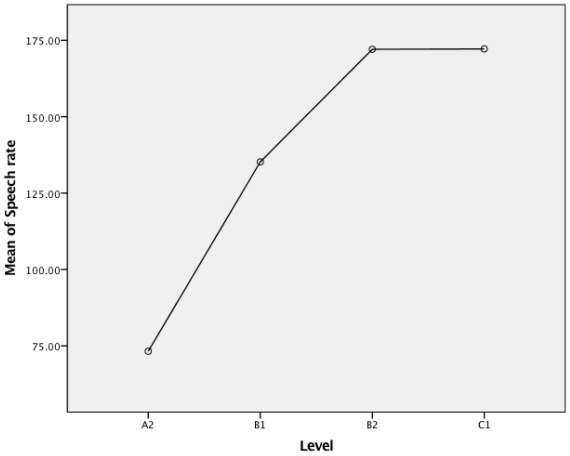
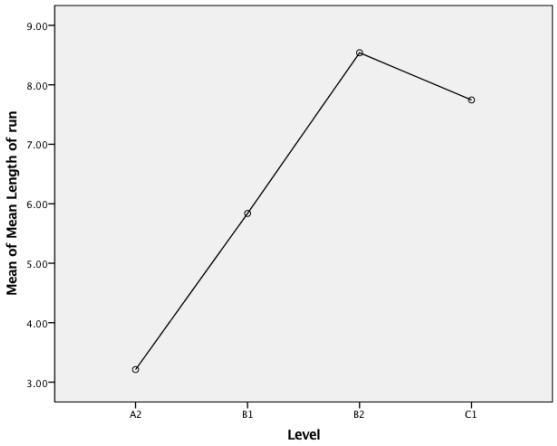


FIGURE 3

Mean Length of Run



Breakdown measures

As noted above, we used a relatively large number of breakdown measures to capture the full picture of how breakdown phenomenon develops across different proficiency levels and tasks.

Silent and filled pauses were examined in terms of their length and frequency and with regard to their location. In what follows, we will first present results for measures of phonation time ratio and length of pauses followed by the results for measures of frequency of pauses. Table 5 shows descriptive statistics for all breakdown measures across levels of proficiency.

TABLE 5: Descriptive Statistics for Breakdown Fluency Measures Across Proficiency Levels

	Phonation time ratio		**Length of all pauses**		**Length of all silent pauses**		**Length of mid-clause silent pauses**		**Length of end-clause silent pauses**		**Length of mid-clause filled pauses**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
A2	46.45	10.32	1.61	.59	1.42	.67	1.25	.69	1.55	.88	.30	.35
B1	71.29	6.59	.75	.17	.63	.23	.54	.18	.68	.30	.47	.29
B2	76.85	5.05	.71	.18	.56	.13	.50	.12	.60	.20	.37	.27
C1	72.91	7.01	.74	.16	.54	.12	.45	.14	.57	.15	.47	.23

	Length of end-clause filled pauses		**Frequency of all pauses**		**Frequency of silent pauses**		**Frequency of filled pauses**		**Frequency of mid-clause silent pauses**		**Frequency of end-clause silent pauses**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
A2	.36	.34	21.40	5.22	29.11	9.87	3.08	3.90	7.47	3.77	7.09	2.96
B1	.39	.31	23.16	3.73	28.63	13.05	7.83	7.83	6.27	3.55	8.05	3.62
B2	.32	.29	19.99	3.68	24.70	10.35	5.29	6.53	3.98	2.39	8.37	3.96
C1	.49	.26	22.00	2.91	23.65	10.58	8.61	5.60	4.34	2.84	7.49	3.41

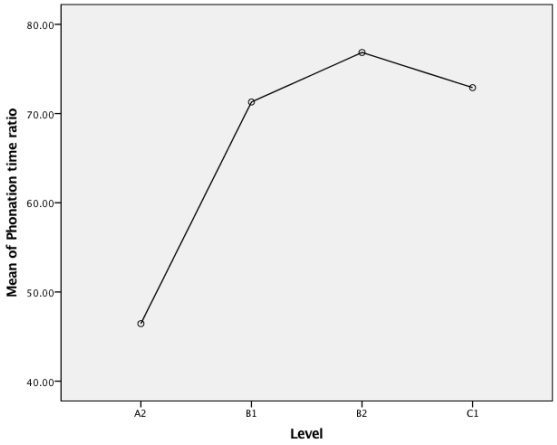
	Frequency of mid-clause filled pauses		**Frequency of end-clause filled pauses**	
	Mean	**SD**	**Mean**	**SD**
A2	.67	1.18	.87	.92
B1	2.36	2.41	1.55	1.96
B2	1.57	2.16	1.07	1.30
C1	2.72	2.07	1.59	1.18

N = 32

As for the phonation time ratio, a significant difference was observed across different levels of proficiency ($F = 93.41$, $p < .001$, $\eta^2 = .710$). The post hoc analysis showed that A2 level was different from all other levels, B1 was different from B2 but not from C1, and B2 and C1 were not different from one another (see Figure 4).

FIGURE 4

Phonation Time Ratio



Length of pauses

As for measures of length of pauses across different levels, a significant difference was observed for both mean length of all pauses (F = 53.84, p < .001, η^2 = .590) and mean length of silent pauses (F = 40.55, p < .001, η^2 = .514). The post-hoc analysis for both measures showed that the A2-level speakers paused significantly longer than B1, B2 and C1 for both measures. B1, B2 and C1 level speakers were not different from one another (see Figures 5 and 6).

FIGURE 5

Mean Length of All Pauses

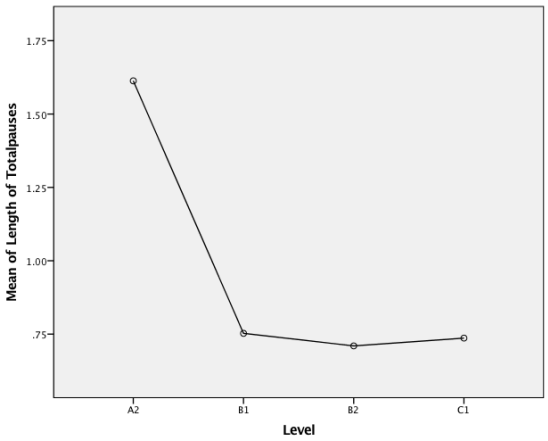
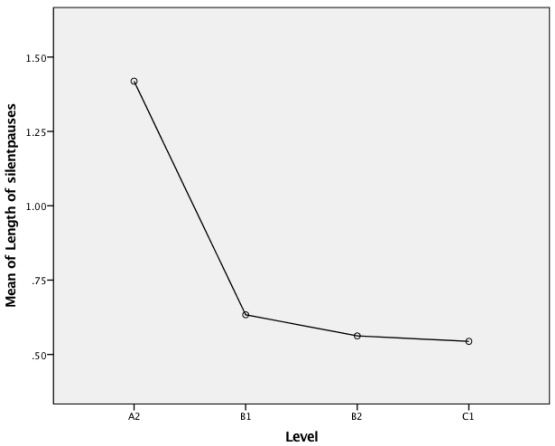


FIGURE 6

Mean Length of Silent Pauses



For mean length of mid-clause silent pauses, a significant difference was observed across different proficiency levels (F = 32.69, p < .001, η^2 = .465). A significant difference was also observed for mean length of end-clause silent pauses across different levels of proficiency (F = 28.71, p < .001, η^2 = .430). The post-hoc analysis for both measures showed that the A2 level was different from B1, B2 and C1. The other levels were not statistically different from one another (Figures 7 and 8).

FIGURE 7

Mean Length of Mid-clause Silent Pauses

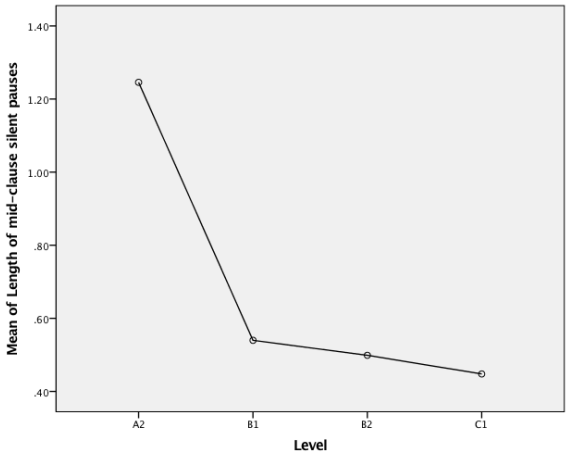
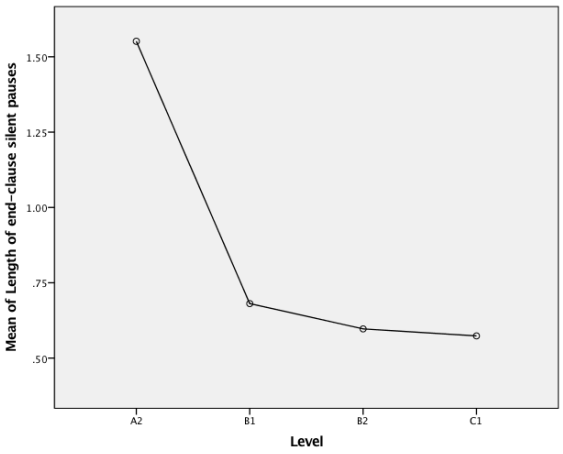


FIGURE 8

Mean Length of End-clause Silent Pauses



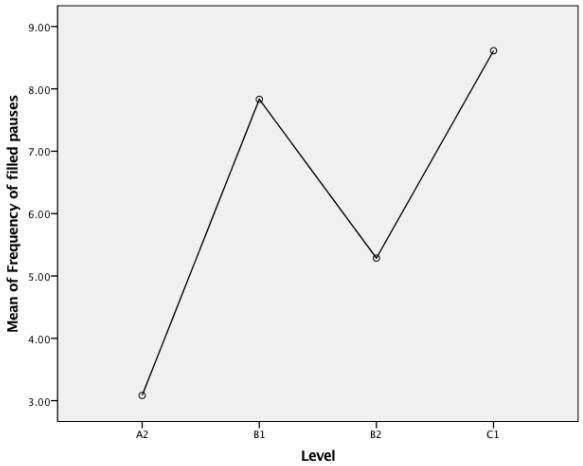
For length of mid-clause and end-clause filled pauses, a significant difference was not observed. Interestingly, the descriptive statistics indicated that the length of mid-clause and end-clause filled pauses seemed to present a non-linear relationship with learners' proficiency levels. C1 and B1 speakers appeared to use longer filled pauses than B2 and A2 speakers.

Frequency of pauses

For frequency of pauses, we used three measures: frequency of all pauses, frequency of silent pauses and frequency of filled pauses. The results of the ANOVAs showed a non-significant difference for frequency of all pauses (F = 3.71, p = .014, η^2 = .09). For frequency of all silent pauses, the results did not indicate a significant difference (p =.149) across levels. Although the results were not significant, the descriptive statistics suggested that there were more silent pauses at lower levels of proficiency (A2 = 29.11, B1= 28.63, B2 = 24.70, C1 = 23.65). Finally, for frequency of filled pauses, a significant difference was observed across different levels of proficiency (F = 4.47, p = .005, η^2 = .103). The post-hoc analysis showed that A2 level was different from B1 and C1, but not from B2. B1, B2 and C1 were not different from each other. The descriptive statistics indicated that with the exception of the B1 level, speakers at higher levels of proficiency produced more filled pauses. Figure 9 shows the results for frequency of filled pauses.

FIGURE 9

Frequency of Filled Pauses

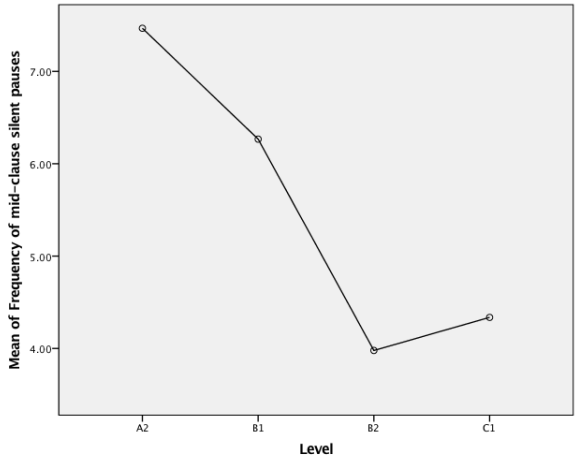


As regards frequency of mid-clause silent pauses, a significant difference was observed across different levels of proficiency (F = 7.67, p < .001, η^2 = .170). The post-hoc analysis showed that A2 and B1 levels were not different from each other, but they were different from

B2 and C1. However, B2 and C1 were not different from each other (see Figure 10). For frequency of end-clause silent pauses, however, the results did not show any significant results across different levels of proficiency (p =.531). The number of end-clause silent pauses at different levels were very similar (A2 = 7.09, B1 = 8.05, B2 = 8.37, C1 = 7.49).

FIGURE 10

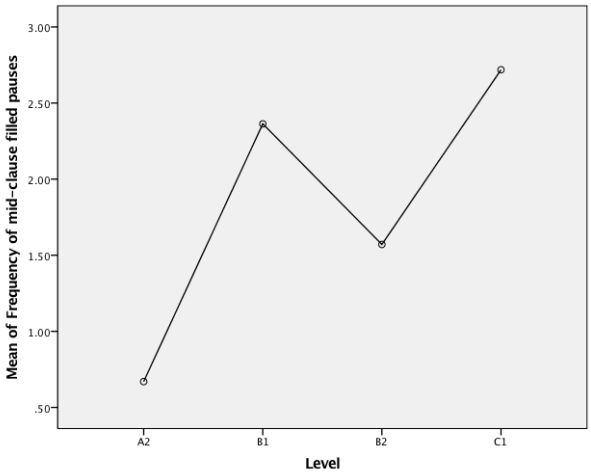
Frequency of Mid-clause Silent Pauses



When we examined frequency of mid-clause filled pauses, a significant difference was observed across different levels of proficiency (F = 5.38, p = .002, η^2 = .120). The post-hoc analysis showed that A2 level was different from B1 and C1, but not from B2. B1, B2 and C1 were not different from each other. Once again, the descriptive statistics implied speakers at C2 and B1 produced more mid-clause filled pauses, suggesting a non-linear relationship between filled pauses and learners' proficiency levels (Figure 11). For frequency of end-clause filled pauses, the results did not show any significant differences across proficiency levels (p = .156). The emerging pattern was however similar, implying C1 and B1 test-takers tended to use more filled pauses (A2 = .87, B1 = 1.55, B2 = 1.07, C1 = 1.59).

FIGURE 11

Frequency of Mid-clause Filled Pauses



Repair measures

Table 6 presents the descriptive statistics for repair measures.

TABLE 6: Descriptive Statistics for Repair Fluency Measures Across Proficiency Levels

	Total repairs		**Reformulations**		**Repetitions**		**Self-corrections**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
A2	4.04	5.89	.83	1.37	2.83	4.50	.38	.65
B1	9.25	5.75	2.78	2.17	5.13	4.45	1.34	1.21
B2	7.63	5.52	2.06	1.79	4.41	3.61	1.16	1.72
C1	8.06	3.73	1.75	1.39	5.09	3.31	1.22	1.58

N = 32

A significant difference was observed for the total number of repairs across different proficiency levels ($F = 4.78$, $p < .004$, $\eta^2 = .110$). The post-hoc analysis showed that A2 level was different from B1 and C1, but not different from B2. The results showed that B1, B2, and C1 levels were not different from one another. It was interesting to see that the B1 level produced the highest and the A2 level the lowest number of repairs (A2 = 4.04, B1 = 9.25, B2 = 7.63, C1 = 8.06). A significant difference was also observed for false starts and

reformulations across different levels of proficiency ($F = 5.95$, $p < .001$, $\eta^2 = .158$). The post-hoc analysis showed that A2 level was different from B1 but not from other levels. There was, however, no significant difference between B1, B2 and C1 levels. Figures 12 and 13 show the results.

FIGURE 12

Total Repair

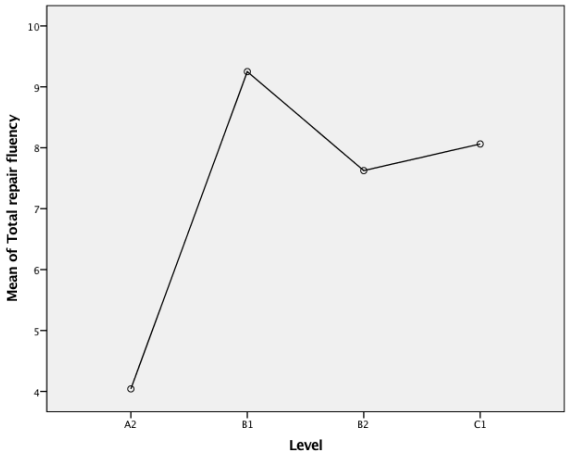
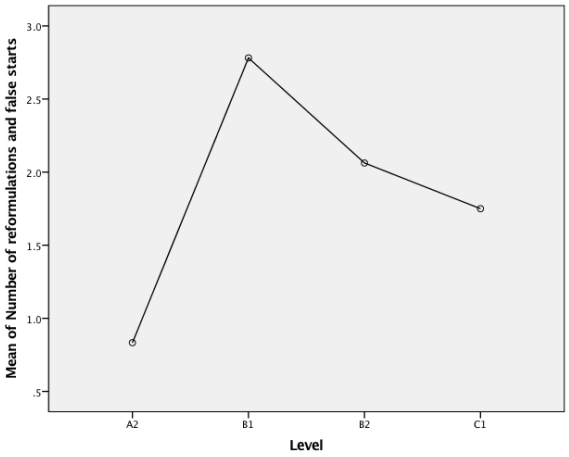


FIGURE 13

False Starts and Reformulations



For number of repetitions, a significant difference was not observed across different levels of proficiency or tasks ($p < .130$). Although statistically non-significant, the results implied that higher proficiency levels produced more repetitions. Neither was a significance difference observed for the number of self-corrections across proficiency levels ($p < .130$).

Table 7 below summarizes the results presented in this section. In this table, the equal signs (=) signify no significant differences, while the arrows (<, >) show that one value was significantly less than or greater than the other and the direction.

TABLE 7: Summary of all ANOVA and Post-hoc Results

Speed and composite measures	**Level**
Articulation rate	(C1=B2)>B1>A2
Speech rate	(C1=B2)>B1>A2
Mean length of run	(C1=B2)>B1>A2

Breakdown measures	**Level**
Phonation time ratio	(B2=C1=B1)>A2, B2>B1
Mean length of all pauses (filled + silent)	A2>(B1=B2=C1)
Mean length of all silent pauses	A2>(B1=B2=C1)
Mean length of mid-clause silent pauses	A2>(B1=B2=C1)
Mean length of end-clause silent pauses	A2>(B1=B2=C1)
Mean length of mid-clause filled pauses	No difference
Mean length of end-clause filled pauses	No difference
Frequency of all pauses (filled + silent)	No difference
Frequency of silent pauses	No difference
Frequency of filled pauses	B1=C1=B2, C1>A2, B1>A2, B2=A2
Frequency of mid-clause silent pauses	(A2=B1)>(C1=B2)
Frequency of end-clause silent pauses	No difference
Frequency of mid-clause filled pauses	C1=B1=B2, C1>A2, B1>A2, B2=A2
Frequency of end-clause filled pauses	No difference
Repair measures	**Level**
Total number of repairs	B1=B2=C1, (B1=C1)>A2, B2=A2
Frequency of false starts & reformulations	B1=B2=C1, B1>A2
Frequency of repetitions	No difference
Frequency of self-corrections	No difference

DISCUSSION

The current study set out to examine which measures best characterize fluency at each level of proficiency assessed in the Aptis Speaking test, and whether any can consistently distinguish one level from the next. The study, in effect, was interested in finding out to what extent different aspects of fluency progress across levels of proficiency in a linear fashion. In the following section, we will first present a summary of the findings, followed by a discussion of the findings, and in the last section the research question of the study will be answered. Effect sizes are also reported for the comparisons, where relevant. Following Plonsky and Oswald (2014), an effect size of .40 is considered small, .70 as medium and 1.00 and above as large.

Speed and composite fluency

Articulation rate, speech rate and mean length of run successfully distinguished performance at different levels of proficiency. The three measures also distinguished A2 and B1 levels from

each other and other levels, with small to moderate effect sizes (Plonsky & Oswald, 2014) ranging from .457 to .710. However, the results indicate that speed and composite fluency is not statistically different between B2 and C1 level. The lack of distinction between B2 and C1 level may demonstrate a ceiling effect, i.e., speed increases with level of proficiency from A2 to, B1 and B2, but not any further. It is also possible to argue that some other measures of fluency (e.g., a smaller pausing threshold) or, indeed, of performance (e.g., lexical sophistication) may be able to distinguish these two levels. Alternatively, it is possible to argue that the result implies that a more demanding task at the C1 level may be necessary in order to distinguish the speed and composite fluency of B2 and C1 levels. The latter speculation, already confirmed in SLA studies investigating task complexity and fluency (e.g., Gilabert et al., 2016; Michel, 2011; Revesz et al., 2016), warrants further investigation.

Breakdown measures: Phonation time ratio and length of pauses

The examination of phonation time ratio indicated a significant difference across different levels of proficiency. A2 level was different from all other levels, B1 was different from B2 but not from C1, and B2 and C1 were not different from one another. Because phonation time ratio captures the extent to which a speaker can fill time with speech, the likely explanation for this finding is that there are differences between proficiency levels in either the *frequency* of silent pauses, the *length* of silent pauses, or both. The analyses of length of pauses were carried out for silent and filled pauses and for pauses at mid-clause and end-clause locations separately. The results suggested that the length of silent pauses distinguishes A2 from other levels of proficiency, while B1, B2 and C1 levels are not different from each other. For length of silent pauses overall (i.e. irrespective of their location), A2 level produced the longest pauses. The effect sizes for these comparisons ranged from .43 to .52, suggesting small but approaching medium-size effects. Interestingly, the differences between length of silent pauses show a pattern of decrease from A2 to C1, demonstrating a linear relationship between

fluency and proficiency level. As for length of filled pauses, none of the comparisons demonstrated a significant difference across proficiency levels. This finding is in line with Segalowitz et al.'s (2017) study in which length of filled pauses did not prove to be a core fluency feature.

Breakdown measures: Frequency of pauses

The analysis of number of pauses revealed several interesting findings. First, unlike the findings for length of silent pauses, analyses of the frequency of silent pauses do not present a consistent pattern. While frequency of mid-clause pauses indicated a significant difference, frequency of all pauses (irrespective of location) and frequency of end-clause pauses did not. Frequency of mid-clause silent pauses distinguished lower levels of proficiency (A2 and B1) from higher levels (B2 and C1), suggesting that at lower levels the speakers produce more silent pauses at mid-clause positions. Overall, frequency of mid-clause silent pauses decreased with an increase in proficiency. This is an interesting finding that consolidates previous research findings (Skehan & Shum, 2017), suggesting that number of mid-clause silent pauses reduces in a linear fashion as proficiency increases. The number of end-clause silent pauses did not distinguish the different levels, implying that speakers' number of pauses at end-clause positions was similar across different proficiency levels. This finding is in line with previous research that claims frequency of mid-clause pausing is a characteristic of L2 speech (Tavakoli, 2011; Derwing et al., 2009) hypothesizing that L2 speech is not automatic, and therefore L2 speakers use the silent pauses to formulate their speech.

For frequency of filled pauses, the results demonstrated that there were statistical differences between the proficiency levels for total number of filled pauses and number of mid-clause filled pauses. In both comparisons, C1 level produced the most and A2 level the least number of filled pauses. Overall, the total number of filled pauses shows a clear and progressive pattern from A2 to B2 and C1, suggesting that candidates at higher levels of proficiency use

filled pauses more frequently. However, B1 speakers did not fit the same pattern, often using as many filled pauses as C1 test-takers. As we will discuss below, B1 level also acts differently on repair measures as they use repairs most frequently. Considering the two patterns together, it is possible to postulate that number of filled pauses and use of repair measures might be interrelated with speakers using a filled pause to activate repair measures. The significant results for number of mid-clause filled pauses indicate that speakers at higher proficiency levels use more filled pauses in general, and more mid-clause filled pauses in particular. It is also interesting to note that the more proficient speakers use mid-clause filled pauses more frequently, whereas less proficient speakers produce more mid-clause silent pauses. One could argue that this use of mid-clause filled pause at higher levels relates to the speaker's communicative strategies, or reflects an awareness of the need to "hold the floor" during speech. To the best of our knowledge, this is the first study examining filled and silent pauses across different proficiency levels, and therefore these findings make a valid contribution to the understanding of breakdown fluency in the field.

Repair measures

The analysis indicated statistically significant differences across proficiency levels in their use of reformulations and total number of repairs. The results showed that B1 level speakers produced the most and A2 level the least number of total repairs, repetitions and reformulations. Kormos (2006) argues that L2 learners' access to declarative knowledge, especially at lower proficiency levels when language use is less automatic, is a key distinction between L1 and L2 production processes. These results may imply that, while engaged in speaking tasks, A2 level speakers may not easily access their declarative knowledge which is necessary for monitoring and repairing speech. Our analysis (e.g., mean length of run and phonation time ratio) of the A2 data showed that these learners were using pauses between

individual words to create utterances, minimizing the need to 'repair' *per se* as the speech was being constructed slowly, word-by-word. The contrast with B1 level is also interesting as B1 level speakers use more repairs while attempting longer runs of speech. It seems to us that at B1 level repair processes are highly activated, and when speakers progress to B2 and C1 level, they use repair measures in moderation. This moderate need for making repairs is inevitably linked with development of L2 in terms of more accuracy and complexity. More research is needed to explore in what ways the development of repair processes interacts with other aspects of speech when proficiency develops. As discussed above, the use of repair measures is also linked to the pausing phenomenon, and therefore any discussion of repair measures should ideally look at the interaction between repair and breakdown aspects of fluency.

Our research question concerned *the extent to which four types of fluency (i.e. speed, breakdown, repair and composite fluency) are presented across different levels of proficiency (CEFR A2, B1, B2, and C1)*. The most important findings are:

1. Speed and composite fluency distinguishes A2, B1 and B2 levels reasonably consistently. B2 and C1 levels are not different in this regard.
2. Length of silent pauses distinguishes A2 level from other proficiency levels. A2 level speakers pause for significantly longer than the other assessed levels.
3. Frequency of mid-clause silent pauses distinguishes lower (A2 and B1) from higher (B2 and C1) proficiency levels. The higher level speakerss' speech is characterized by fewer mid-clause silences.
4. Frequency of filled pauses distinguishes A2 from higher levels.
5. Higher proficiency levels generally use filled pauses more frequently than lower levels, but the excessive use of filled pauses by B1 speakers makes the progression non-linear.

6. Repair measures (both total number and false starts & reformulations) distinguish A2 and B1 levels as the former produces very few and the latter most repairs. While B2 and C1 levels engage in repairs to a moderate degree, B1 level actively uses repair measures to reformulate speech.

It is necessary to note that in the absence of any L1 data from the participants, it is difficult to claim in full certainty that the results obtained here are only due to the speakers' L2 fluency behaviour and not affected by the individual differences in their L1 speaking style.

CONCLUSIONS

The findings of the current study have important implications for SLA research as this is the first study examining fluency across assessed levels of proficiency in a detailed manner carefully operationalizing a wide range of micro-analysis of fluency features. The results suggest that speed and composite measures consistently distinguish fluency across proficiency levels, length of silent pauses is a key characteristic of low proficiency performance (A2), and frequency of silent pauses distinguishes the lower (A2 and B1) from higher proficiency levels. Notwithstanding the significance of such findings for fluency research and measurement, and in line with other researchers in the field (de Jong, 2018; Housen et al.2012; Kormos, 2006; Skehan, 2009, 2015; Tavakoli & Hunter, 2018), we argue that adopting a more finely-grained approach to analysing fluency can not only help develop a better understanding of how L2 is processed and produced, but it can contribute towards the development of an L2 speech production model. In this regard, the current study has revealed interesting differences between the fluency construct at different proficiency levels, implying that there may be different processes at work when speakers of different proficiency level produce L2. One may argue that at lower proficiency levels, speakers rely more predominantly on longer silent pauses to process and produce speech, whereas more proficient speakers use filled pauses and repair to punctuate and structure speech. The finding that A2 level speakers did not produce many

repairs or filled pauses is important for SLA research as it may suggest a threshold proficiency level (length of run) is required for the repair processes to be activated. A key methodological implication of the findings of the study for SLA research is that non-linear developmental trajectories may be observed when more nuanced fluency measures are employed.

A greater understanding of how different aspects of fluency are represented at different levels of proficiency can also help with the development and validation of more precise, fine-grained descriptors for language testing. This study offers evidence-based findings that are central to a systematic approach to constructing and rewording rating scales used in the assessment of fluency. Given the high-stake nature of these tests and the far-reaching impact they have on test-takers' employment and education, it seems crucial that language testing organizations consider such findings in their rating scale development and revisions.

The findings should also be considered in rater training programmes as they can help raters develop a more in-depth understanding of fluency features across proficiency levels. While further studies are necessary to confirm whether the features identified in this study are actually salient to raters when rating spoken performance in real time (e.g., Brown, 2006b; Brown et al., 2005; Ducasse & Brown, 2009; May, 2011; Orr, 2002; Pollitt and Murray, 1996), it is hoped that the comprehensive analysis of fluency characteristics in this study enables test providers to make an informed decision about constructing fluency rating descriptors and training raters. Last but not least, the findings of the current study have important implications for L2 teachers and teaching. These findings can offer teachers with a better understanding of how fluency patterns vary at different levels of proficiency, what characterizes learner fluency at each level, and what can be done to help learners to improve their fluency at different levels of proficiency.

NOTES

[i] Although terms such as ‘adequate speed’, ‘more ease’ and ‘less hesitation’ are often used to give an objective dimension to defining fluency, they naturally represent subjective notions of what is perceived as ‘adequate’ or ‘more ease’. Despite their importance in understanding and defining fluency, these terms have not been operationalized in fluency research yet. If a meaningful and objective measurement of fluency is expected, reference measures for the terms should be established.

[ii] Articulation Rate and Syllable Length are calculated in a similar manner. Articulation rate divides the number of syllables produced by the phonation time while syllable length divides phonation time by the number of syllables produced.

[iii] The total amount of time for the samples varied for each task in line with the requirements of the Aptis speaking test (see Table 2). Phonation time was calculated as the time from when the test-taker began to speak until after the last syllable of speech produced. This means that there were small differences in the phonation time between subjects even on the same task. Multiplication by 60 is performed to provide a ‘per minute’ measure.

[iv] We also ran Discriminant Function Analysis (DFA) to cross-examine our results. Although the findings of DFA confirmed the results of the ANOVAs, we have found ANOVAs more suitable in demonstrating the changes in utterance fluency across different proficiency levels.

[v] We consider phonation time ratio a breakdown measure as it indicates what proportions of one’s speech are phonation and silence. Therefore, a person who pauses either frequently or for long periods will have a lower phonation time ratio than someone who pauses infrequently or only for short periods.

REFERENCES

Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Variables affecting L2 gains during study abroad. *Foreign Language Annals*, *47*, 464-486.

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer [Computer program] Version 5.3.5.1, retrieved from http://www.praat.org/

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

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*, 886–899.

Brown, A. (2006a). An examination of the rating process in the revised IELTS Speaking Test. In P. McGovern, & S. Walsh (Eds.), *IELTS research reports* 2006 (pp. 41-70), Canberra & Manchester: IELTS Australia and British Council.

Brown, A. (2006b). Candidate discourse in the revised IELTS Speaking Test. In P. McGovern, & S. Walsh (Eds.), *IELTS research reports 2006* (pp. 71-89), Canberra & Manchester: IELTS Australia and British Council.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test taker performance on English-for-Academic-Purposes speaking tasks* (TOEFL Monograph No. 29). Princeton, NJ: Educational Testing Service.

Clark, H. H., & Fox Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*, 73-111.

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

de Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly, 15*, 237-254.

de Jong, N. H. (2016a). Predicting pauses in L1 and L2 speech: the effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*, 113-132.

de Jong, N.H. (2016b), Fluency in second language assessment. In D. Tsagari and J. Banerjee (Eds.). *Handbook of Second Language Assessment* (pp.203-218). Amsterdam: Mouton de Gruyter.

de Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In R. Eklund (Ed.), *Proceedings of the 6th workshop on disfluency in spontaneous speech (DiSS)* (pp. 17-20), Stockholm: Royal Institute of Technology (KTH).

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 behaviour. *Applied Psycholinguistics*, *36*, 223-243.

de Jong, N. H., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*, 533-568.

de Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2012). The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (p.121-142), Amsterdam: John Benjamins

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*, 533–557.

Dewaele, J-M. (1996). How to measure formality of speech? A Model of Synchronic Variation. In K. Sajavaara & C. Fairweather (eds.), *Approaches to second language acquisition, Jyväskylä Cross-Language Studies* 17 (pp. 119-133), Jyväskylä: University of Jyväskylä.

Dörnyei, Z., & Kormos, J. (1998). Problem-solving mechanisms in L2 communication. *Studies in second language acquisition, 20*, 349-385.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language testing, 26*, 423-443.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 65‑111). Cambridge: Cambridge University Press.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21, 354-75.

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Longman/Pearson Education.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*, 208-238.

Gilabert, R., Manchon, R., & Vasylets, O. (2016). Mode in theoretical and empirical TBLT research: Advancing research agendas. *Annual Review of Applied Linguistics, 36*, 117–135.

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*, 379-399.

Hilton, H. (2014). Oral fluency and spoken proficiency: considerations for research and testing. In P. Leclerq, A. Edmands & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 27-51). Bristol: Multilingual Matters.

Housen, A., Kuiken, F. & Vedder, I. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol.32). Amsterdam: John Benjamins.

Huensch, A., & Tracy-Ventura, N. (2016). Understanding second language fluency behaviour: The effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycholinguistics, 38*, 755-785.

Huensch, A., & Tracy–Ventura, N. (2017). L2 Utterance Fluency Development Before, During, and After Residence Abroad: A Multidimensional Investigation. *The Modern Language Journal*, *101*, 275-293.

Hunter, A-M. (2017). Fluency development in the ESL classroom: The impact of immediate task repetition and procedural repetition on learners' oral fluency. Unpublished doctoral dissertation, University of Surrey, Guildford, UK.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*:24-49.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning, 64*, 809-854.

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

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*, 145-164.

Lennon, P. (1990). Investigating fluency in EFL: a quantitative approach. *Language Learning, 40*(3), 387-417.

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

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

May, L. (2011). *Interaction in a Paired Speaking Test*, Frankfurt am Main: Peter Lang.

Michel, M. (2011). Effects of task complexity and interaction on L2 performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 141–174). Amsterdam: John Benjamins.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, *30*, 143-154.

O'Sullivan, B., & Dunlea, J. (2015). *Aptis General Technical Manual Version 1.0*. Technical Report TR/2015/005. British Council: London.

Nakatsuhara, F. (2014). *A research report on the development of the Test of English for Academic Purposes (TEAP) speaking paper for Japanese University entrants.* Eiken Foundation of Japan. Available online at: https://www.eiken.or.jp/teap/group/pdf/teap_speaking_report1.pdf

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly*, *14*, 1-18.

Peltonen, P. (2018). Exploring Connections Between First and Second Language Fluency: A Mixed Methods Approach. *The Modern Language Journal*, *102*, 676-692.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

Pollitt, A., & Murray, N. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance, testing, cognition and assessment* (pp. 74-91). Cambridge: UCLES/Cambridge University Press.

Préfontaine, Y. (2013). Perceptions of French fluency in second language speech production. *Canadian Modern Language Review, 69*, 324-348.

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

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*, 828-848.

Schmid, M. S., & Fägersten, K. B. (2010). Disfluency markers in L1 attrition. *Language learning, 60*, 753-791.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York: 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, 90-116.

Skehan, P. (2003). Task-based instruction. *Language Teaching, 36*, 1–14.

Skehan, P. (2009). Modelling Second Language Performance: Integrating Complexity, Accuracy, Fluency, and Lexis. *Applied Linguistics, 30*, 510-532.

Skehan, P. (2014). Limited attentional capacity, second language performance, and task-based pedagogy. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 211–260). Amsterdam, the Netherlands: John Benjamins.

Skehan, P. (2015). Limited Attention Capacity and Cognition: Two hypotheses regarding second language performance on tasks. In M. Bygate (Ed.), Domains and directions

in the development of TBLT: A decade of plenaries from the international conference (pp. 123-156). Amsterdam: John Benjamins.

Skehan, P. & Foster, P. (2008). Complexity, accuracy, fluency and lexis in task-based performance: A meta-analysis of the Ealing research. In Van Daele, S., Housen, A., Kuiken, F., Pierrard, M. & Vedder, I. (Eds.). *Complexity, Accuracy, and Fluency in Second Language Use, Learning, and Teaching* (pp. 207-226). Brussels: Contactforum.

Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics in Language Teaching*, *54*, 97-111.

Skehan, P. & Shum, S. (2017). What influences performance? Personal style or the task being done? In L. Wong & K. Hyland (Eds.), *Faces of English education: Students, teachers and pedagogy* (pp. 29-43). London: Taylor and Francis.

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers, *ELT Journal, 65*, 71–79.

Tavakoli, P. & Hunter, A-M. Is fluency being 'neglected' in the classroom? Teacher understanding of fluency and related classroom practices. *Language Teaching Research, 22*, 330-349.

Tavakoli, P. & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-277). Amsterdam: Benjamins.

Tavakoli, P. Nakatsuhara, F. & Hunter, A-M. *Scoring validity of the Aptis Speaking test: Investigating fluency across tasks and levels of proficiency.* ARAGs Research Reports Online. ISSN 2057-5203London: British Council.

Witton-Davies, G. (2014). *The study of fluency and its development in monologue and dialogue*. Unpublished doctoral dissertation, Lancaster University, Lancaster, UK.