Scoring validity of the Aptis Speaking test: investigating fluency across tasks and levels of proficiency

Article

Published Version

BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

SCORING VALIDITY OF THE APTIS SPEAKING TEST: INVESTIGATING FLUENCY ACROSS TASKS AND LEVELS OF PROFICIENCY

AR-G/2017/7

Parvaneh Tavakoli, University of Reading
Fumiyo Nakatsuhara, University of Bedfordshire
Ann-Marie Hunter, St. Mary's University

ARAGs RESEARCH REPORTS ONLINE
SERIES EDITOR: VIVIEN BERRY
ISSN 2057-5203

ABSTRACT

Second language oral fluency has long been considered as an important construct in communicative language ability (e.g. de Jong et al, 2012) and many speaking tests are designed to measure fluency aspect(s) of candidates' language (e.g. IELTS, TOEFL iBT, PTE Academic). Current research in second language acquisition suggests that a number of measures of speed, breakdown and repair fluency can reliably assess fluency and predict proficiency. However, there is little research evidence to indicate which measures best characterise fluency at each level of proficiency, and which can consistently distinguish one proficiency level from the next. This study is an attempt to help answer these questions.

This study investigated fluency constructs across four different levels of proficiency (A2–C1) and four different semi-direct speaking test tasks performed by 32 candidates taking the Aptis Speaking test. Using PRAAT (Boersma & Weenik, 2013), we analysed 120 task performances on different aspects of utterance fluency including speed, breakdown and repair measures across different tasks and levels of proficiency. The results suggest that speed measures consistently distinguish fluency across different levels of proficiency, and many of the breakdown measures differentiate between lower (A2, B1) and higher levels (B2, C1). The varied use of repair measures at different proficiency levels and tasks suggest that a more complex process is at play. The non-significant differences between most of fluency measures in the four tasks suggest that fluency is not affected by task type in the Aptis Speaking test. The implications of the findings are discussed in relation to the Aptis Speaking test fluency rating scales and rater training materials.

Authors

Dr. Parvaneh Tavakoli is an Associate Professor in Second Language Acquisition and Applied Linguistics at the University of Reading. She holds a PhD in Applied Linguistics and an MA in TEFL. Parvaneh is a Senior Fellow of Higher Education Academy in the UK. She has been teaching in the UK and abroad since 1991, and has led a number of national and international research projects. Her research interests include second language acquisition, language testing and assessment, task-based language teaching and teacher training.

Dr. Fumiyo Nakatsuhara is a Reader at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. Her research interests include the nature of co-constructed interaction in various speaking test formats (e.g. interview, paired and group formats), task design and rating scale development. Fumiyo's publications include the book, *The Co-construction of Conversation in Group Oral Tests* (2013, Peter Lang), book chapters in *Language Testing: Theories and Practices* (O'Sullivan, ed. 2011) and *IELTS Collected Papers 2: Research in Reading and Listening Assessment* (Taylor and Weir, eds. 2012), as well as journal articles in *Language Testing* (2011; 2014) and *Language Assessment Quarterly* (2017). She has carried out a number of international testing projects, working with ministries, universities and examination boards.

Dr. Ann-Marie Hunter was awarded her PhD from St Mary's University and University of Surrey in August 2017. Her research focuses on second language speech performance and task-based approaches to language teaching, and seeks to identify ways that language teachers can work with the construct of oral fluency in a language classroom. She is currently based in Manchester and is a visiting lecturer in TESOL and Second Language Acquisition at the University of Leeds.

CONTENTS

1. INTRODUCTION 5

2. THEORETICAL BACKGROUND 5

3. FLUENCY IN LANGUAGE TESTING RESEARCH 7

4. MEASURING FLUENCY 9

5. RESEARCH AIMS AND QUESTIONS 11

6. ASSESSMENT OF SPEAKING IN APTIS AND RATING OF FLUENCY 12

7. METHODOLOGY 13

7.1. Research design 13
7.2. Data set 13
7.3. Analytic fluency measures 14
7.4. Data analysis procedures 14
7.5. Using PRAAT 15

8. ANALYSES AND RESULTS 16

8.1. Repeated-measures MANOVA 17
8.2. Univariate analysis 17
8.3. Fluency measures across proficiency levels and tasks 17
8.3.1. Speed measures 18
8.3.2. Breakdown measures 20
8.3.3. Repair measures 26

9. DISCUSSION OF THE FINDINGS 29

10. RECOMMENDATIONS 34

10.1 Recommendations for the Aptis Speaking rating scales 34
10.2 Recommendations for the Aptis Speaking training materials 36
10.2.1 Analysis of the Aptis Speaking training materials 36
10.2.2 Recommendations 37

11. CONCLUSIONS AND WAYS FORWARD 38

REFERENCES 41

APPENDIX 1: Aptis Speaking rating scales 44

APPENDIX 2: Descriptive statistics 47

APPENDIX 3: Glossary 54

List of tables

Table 1: Fluency-related rating descriptors in selected standardised tests ... 8
Table 2: Structure of the Aptis Speaking Test ... 12
Table 3: Fluency descriptors across tasks and proficiency levels in Aptis ... 12
Table 4: Summary of the level and task comparisons of all analytic measures ... 31
Table 5: Cognitive processing model of speaking ability (Field, 2011: 74–77) ... 33
Table 6: Suggested fluency descriptors for Task 1 ... 34
Table 7: Suggested fluency descriptors for Tasks 2 and 3 ... 35
Table 8: Suggested fluency descriptors for Task 4 ... 35

Table A2.1: Descriptive statistics for fluency measures across proficiency levels ... 47
Table A2.2: Descriptive statistics for fluency measures across tasks ... 50

List of figures

Figure 1: Speech rate across levels and tasks ... 18
Figure 2: Articulation rate across levels and tasks ... 19
Figure 3: Mean length of run across levels and tasks ... 19
Figure 4: Phonation time ratio across levels and tasks ... 20
Figure 5: Total length of pauses across levels and tasks ... 21
Figure 6: Length of mid-clause silent pauses across levels and tasks ... 21
Figure 7: Length of end-clause silent pauses across levels and tasks ... 22
Figure 8: Length of mid-clause filled pauses across levels and tasks ... 22
Figure 9: Length of end-clause filled pauses across levels and tasks ... 23
Figure 10: Number of silent pauses across levels and tasks ... 23
Figure 11: Number of mid-clause silent pauses across levels and tasks ... 24
Figure 12: Number of end-clause silent pauses across levels and tasks ... 24
Figure 13: Total number of filled pauses across levels and tasks ... 25
Figure 14: Number of mid-clause filled pauses across levels and tasks ... 25
Figure 15: Number of end-clause filled pauses ... 26
Figure 16: Total number of repair measures across levels and tasks ... 26
Figure 17: Number of false starts and reformulations across levels and tasks ... 27
Figure 18: Number of repetitions across levels of proficiency and tasks ... 28
Figure 19: Mean number of self-corrections ... 28

1. INTRODUCTION

Fluency has long been recognised as a key characteristic of spoken language ability, a major component of the construct of speaking (Fulcher, 2003), and a descriptor of spoken proficiency in several widely-accepted language benchmarks (e.g. FSI, 1970s, ACTEFL, 2014 and CEFR, 2001). Fluency is also featured in rating scales in most standardised speaking tests, such as Aptis, Cambridge General English Tests, IELTS, PTE Academic and TOEFL. Despite the significant role it plays in the assessment of second language speaking ability, fluency is usually represented in a rather limited way, with only few of its fundamental features presented in rating scales. While an evidence-based approach to rating scale development has gained currency over the past decade (e.g. Brown et al., 2005; Fulcher, 1996; Nakatsuhara, 2014), what characteristics of fluent speech are relevant to differentiate levels of proficiency and can, therefore, be used as useful criterial features are still relatively under-researched. This is the gap that the current study aims to help fill.

The overall aim of this project is to contribute to the British Council's research agenda on test validation by investigating criterial performance features in speaking at the CEFR levels assessed in Aptis. The specific aim of the project is to examine what characteristics of fluent speech are relevant across different levels of proficiency (A2–C1), what impact task design may have on speech fluency at different levels, and how various aspects of fluency can be effectively deployed in the Aptis operational rating scales. The research findings are expected to help enhance the scoring validity of Aptis Speaking, by offering a better understanding of its fluency constructs and by providing fluency performance benchmarks for the A2–C1 levels which the test was designed to measure.

2. THEORETICAL BACKGROUND

The important role of fluency in communicative language ability has been repeatedly highlighted in second language acquisition research (Segalowitz, 2000, 2010). A growing interest is also observed in the number of studies exploring the relationship between fluency and communicative adequacy (De Jong et al., 2012; Revesz et al., 2016) and highlighting its significance in second language teaching and learning (Mora & Valls-Ferrer, 2012). In the field of second language pedagogy, this interest has led to a number of changes at a language teaching policy level, including the introduction of fluency in the national curriculum for languages (e.g. UK GCSE curriculum, January 2015).

In English language learning, fluency can represent two different but interrelated concepts. In its broader sense, fluency refers to a speaker's overall speaking proficiency and it may refer to his/her skills in use of language for communication purposes effectively. In its technical sense, fluency refers to ease or automaticity with which speech is produced, often demonstrated through flow, continuity and smoothness of speech (Segalowitz, 2010; Skehan, 2014). Researchers have argued that L2 speech fluency is a complex and multifaceted construct that covers a multitude of different sub-components, e.g. linguistic, psycholinguistic and sociolinguistics factors, potentially interacting with one another during the speech production process (Kormos, 2006; Lennon, 2000; Segalowitz, 2000, 2010). Highlighting the multifaceted nature of fluency, previous research has concluded that fluency is a difficult construct to define (Freed, 2000) and a complex performance feature to measure (De Jong et al., 2011; Witton-Davies, 2014).

While the current conceptualisation of fluency is still limited (Kahng, 2014), recent studies in second language acquisition have shed light on the nature of fluency and have offered a more systematic and evidence-based approach to defining fluency. Segalowitz (2010), for example, proposes that L2 speech fluency has three distinct but inter-related aspects: cognitive, utterance and perceived fluency. Cognitive fluency, in Segalowitz's framework, focuses on "the efficiency of the operation of the cognitive mechanisms underlying performance" (Segalowitz, 2000: 202) and "the ability to efficiently mobilize and integrate the underlying cognitive processes responsible for producing utterances" (Segalowitz, 2010: 48); utterance fluency is concerned with the measurable aspects of fluency such as speed, pausing and hesitation; and perceived fluency highlights the inferences listeners make about someone's cognitive fluency based on their perceptions of fluent speech, i.e. the measurable aspects of the speakers' fluency. Segalowitz (2010, 2016) argues that while the three aspects interact with one another, utterance fluency is the most readily measurable aspect of fluency. For the purpose of the current project, we aim to focus on what Segalowitz considers utterance fluency.

In addition to L2 processing demands, several other internal and external factors influence speech fluency. Research in this area has shown that personal speaking style (Derwing et al., 2009; De Jong et al., 2015), L1 typology and cultural norms (De Jong et al., 2015), task design (Foster & Tavakoli, 2009), conditions under which task is performed (Ahmadian, 2012), and social and psychological features of the speech act (Segalowitz, 2016) are some of the factors that affect L2 fluency.

Considering the effects of task design on (L2) performance, L2 acquisition research has provided ample evidence that fluency is task dependent, and therefore, factors such as task design and discourse type affect fluency in significant ways (Michel, 2011; Robinson, 2007; Tavakoli, 2016). Task design features shown to have an impact on language performance include task structure (Tavakoli & Skehan, 2005), storyline complexity, (Tavakoli & Foster, 2008), immediacy of information (Gilabert, 2007), and intentional reasoning (Ishikawa, 2008). Given the reported impact of task design on L2 fluency in language teaching and use contexts, it seems crucial to investigate and understand the effects of task design on fluency under testing conditions. The findings of such research will be of significance for the development and validation of rating scales, rater training programs and test design and development; they will enable test designers/providers to make an informed decision about the choice of test tasks when developing and validating elicitation tasks and fluency rating descriptors.

Findings of recent studies investigating the relationship between speech fluency and communicative adequacy are also of relevance to our study. Examining the componential nature of L2 ability, De Jong et al. (2015) report that while vocabulary is the strongest predictor of language proficiency, aspects of fluency, e.g. speed of performance, are strongly associated with speaking proficiency. Examining performance of 100 L2 users across five different tasks, Revesz et al. (2016), report that some fluency measures, such as speed fluency and pause frequency, are reliable predictors of communicative adequacy, and frequency of filled pause is the strongest predictor of communicative language ability. Kahng (2014) also demonstrates that speech rate and mean length of run are strongly associated with oral proficiency. These findings confirm the direct relationship between fluency and overall language proficiency, and highlight the need for a careful examination of fluency in language testing contexts. As such, conducting research in this area would undoubtedly enable us to understand better how different fluency aspects can be deployed as useful criterial features of learners' language proficiency, thus helping to develop a more reliable operationalisation of fluency in speaking rating scales.

3. FLUENCY IN LANGUAGE TESTING RESEARCH

In language assessment, fluency has long been recognised as a prime characteristic of spoken language ability (e.g. FSI scales in 1970s) and defined as a key descriptor of spoken proficiency in many language level benchmarks (e.g. CEFR, 2001). Indeed, fluency is one of the most common features referred to in both holistic and analytic rating scales used in standardised speaking tests (e.g. Aptis, Cambridge General English Tests, IELTS, PTE Academic, TOEFL). However, the lack of consistency and rigour in the measurement of fluency in speaking assessment has been criticised in recent literature (Kormos, 2006; Tavakoli, 2016). Our informal examination of publicly available rating descriptors in those language benchmarks and standardised tests also suggests that fluency is usually represented in a rather limited or ambiguous way, with only few of its fundamental aspects presented in rating scales.

Table 1 exemplifies fluency-related rating descriptors used in selected standardised tests. As shown in Table 1, some tests such as IELTS and PTE Academic have rather lengthy descriptors on fluency, while others have relatively short descriptors. (It should be noted that the Speaking test of PTE Academic is machine-scored, so the descriptors are not used for actual rating purposes.) 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, and evenness of speech. Some tests also refer to the underlying cause of the hesitation, for example, whether it is content-related or language-related.

However, it is not simply the case that the longer and more detailed the descriptors of fluency, the better. Indeed, what is crucial in operational tests is to provide raters with descriptors that are *useful* (e.g. Taylor & Galaczi, 2011). In other words, descriptors have to be concise and succinct and also include the necessary details to guide raters in making their judgements on fluency. Test designers, therefore, need to strike an optimal balance between construct coverage and rater-usability.

Test (CEFR level, if level-specific)	Cambridge First (B2)		IELTS		PTE Academic		TOEFL iBT		Trinity ISE II (B2)	
Rating category in which fluency is featured	**Discourse management**		**Fluency and coherence**		**Oral fluency**		**Delivery**		**Delivery**	
Fluency-related descriptors	5	• Produces extended stretches of language with very little hesitation.	9	• Speaks fluently with only rare repetition or self-correction; any hesitation is content-related rather than to find words or grammar.	5	• Speech shows smooth rhythm and phrasing. There are no hesitations, repetitions, false starts of non-native phonological simplifications.	4	• Generally well-paced flow (fluid expression).	4	• Speaks promptly and fluently.
	3	• Produces extended stretches of language despite some hesitation.	8	• Speaks fluently with only occasional repetition or self-correction; hesitation is usually content-related and only rarely to search for language.	4	• Speech has an acceptable rhythm with appropriate phrasing and word emphasis. There is no more than one hesitation, one repetition or a false start. There are no significant non-native phonological simplifications.	3	• Speech is generally clear, with some fluidity of expression, though minor difficulties or pausing are noticeable.	3	• Speaks promptly and fluently – occasionally affected by some hesitancy.
	1	• Produces responses which are extended beyond short phrases, despite hesitation.	7	• Speaks at length without noticeable effort; or may demonstrate language-related hesitation at times, or some repetition and/or self-correction.	3	• Speech is at an acceptable speed but may be uneven. There are few repetitions or false starts. There are no long pauses and speech does not sound staccato.	2	• Speech is basically intelligible, though listener effort is needed because of choppy rhythm/pace.	2	• Speaks promptly and fluently enough to follow.
			6	• Is willing to speak at length, though may lose coherence at times due to occasional repetition, self-correction or hesitation.	2	• Speech may be uneven or staccato. Speech (if >=6 words) has at least one smooth three-word run and no more than two or three hesitations, repetitions or false starts. There may be one long pause, but not two or more.	1	• Delivery is choppy, fragmented or telegraphic; frequent pauses and hesitations.	1	• Speaks slowly, sometimes or often halted by hesitancy
			5	• Usually maintains flow of speech but uses repetition, self-correction or hesitation.	1	• Speech has irregular phrasing or sentence rhythm. Poor phrasing, staccato or syllabic timing, and/or multiple hesitations, repetitions, and/or false starts make spoken performance notably uneven or discontinuous. Long utterances may have one or two long pauses and inappropriate sentence-level word emphasis.				
			4	• Cannot respond without noticeable pauses and may speak slowly, with frequent repetition and self-correction.	0	• Speech is slow and laboured with little discernible phrase grouping, multiple hesitations, pauses, false starts, and/or major phonological simplifications. Most words are isolated, and there may be more than one long pause.				
			3	• Speaks with long pauses						
			2	• Pauses lengthily before most words.						

Table 1: Fluency-related rating descriptors in selected standardised tests

As noted earlier, the importance of an evidence-based approach to rating scale development and validation has long been advocated (e.g. Brown, 2006a; Brown et al., 2005, Fulcher, 1996; Nakatsuhara, 2014). In Brown, et al.'s (2005) large-scale validation study on TOEFL, they analysed 198 speech samples taken from test-takers of five proficiency levels. Among different analytic measures, various fluency aspects were examined. ANOVAs were performed with the number of filled pauses per 60 seconds, the number of unfilled pauses per 60 seconds, total pause time, the number of repairs per 60 seconds, speech rate, and mean length of run as dependent variables. Significant differences were found for speech rate, unfilled pauses, and total pause time, with medium or small effect sizes.

More recently, a similar method was used when new rating scales for the TEAP (Test of English for Academic Purposes) Speaking test were developed (Nakatsuhara, 2014). In the process of verifying or suggesting modifications to draft rating scales, a small number of speech samples (N = 23) from a pilot test were analysed with various analytic measures that correspond to draft rating scales. For fluency, the number of unfilled pauses per 50 words, total pause time as a percentage of speaking time, the ratio of repair, false starts and repetition to AS-units, speech rate and articulation rate were compared across three proficiency groups rated by the draft fluency rating scale. Although the small sample size of the study did not allow the use of inferential statistics, the means of the three proficiency groups on all fluency measures varied in accordance with the rating scores that the pilot test-takers obtained. As such, the linguistic analysis confirmed the usefulness of the draft rating descriptors.

While the findings of this body of research have proved useful for rating scale development and validation, our knowledge of what characteristics of fluent speech are relevant across different levels of proficiency, and what impact task design may have on fluency at different levels is still relatively limited, and the area is largely under-researched. This is the gap that the current study aims to help fill. In addition, the focus on the fluency criterion is of particular importance, since research (e.g. Brown, 2006b; Nakatsuhara, 2012) has shown that examiners often find fluency the most difficult to assess.

4. MEASURING FLUENCY

A key question the current study aims to answer is to identify measures of fluency that most consistently represent the construct of L2 speech fluency across different tasks and proficiency levels. Before discussing the measurement of utterance fluency, a brief historical background on measuring fluency is provided.

In an attempt to create a more coherent approach to measuring fluency, Skehan (2003), and Tavakoli and Skehan (2005) called for a more systematic measurement of fluency that represented three key characteristics of fluency: a) speed fluency, i.e. speed with which speech is produced, b) breakdown fluency, i.e. the pauses and silences that break down the flow of speech, and c) repair fluency, i.e. hesitations, repetitions and reformulations that are used to repair speech during the production process. In line with this framework, research in second language acquisition (SLA) has now developed a more detailed and systematic approach to measuring fluency. For measuring speed fluency, a number of major changes have occurred in the operationalisation and measurement of speed. First, there is a more in-depth understanding of speed and a more reliable awareness of the relationship between speed fluency and the pausing phenomenon in speech. Researchers are now aware that speed should be calculated both independently of pauses, e.g. articulation rate, or in a composite form where it is combined with pauses, i.e. speech rate, and that these two measures provide very useful information about fluency.

Another major shift in the measurement of speed fluency has been the introduction of digital technology that has allowed for a more careful and effective way of measuring speed and pauses. While there are several programs that help researchers measure speed accurately (e.g. Audacity and GoldWave Digital Editor), PRAAT (Boersma & Weenik, 2013), a computer program that allows for analysing, synthesising and manipulating speech, is increasingly popular in fluency studies as it allows sounds and files to be annotated in considerable detail and scripts to be written for special commands (see Sections 7.4 and 7.5 below).

In terms of the breakdown, or pause-related, aspect of fluency, there have been developments in how pauses are perceived and operationalised in terms of length, quality and location. Whereas research in the 1990s and early 2000s considered a 1-second pause (Foster & Skehan, 1996) or a 0.4 second pause (Freed, 2000) as a noticeable silence, recent research (De Jong et al., 2012; De Jong & Bosker, 2013) has indicated that for native speakers of English, a pause of longer than 0.25 of a second is considered to be a noticeable pause. More recently, several studies have provided evidence about the importance of the *location* of pauses, i.e. pauses occurring in mid-clause or end-clause positions. Different researchers have suggested that while the frequency of pauses is an important factor that affects listeners' perceptions of fluency, the location of pauses might have an even greater impact, i.e. mid-clause pauses have a more detrimental impact on fluency than end-clause pauses. Tavakoli (2011), for example, showed that what distinguished L2 speakers of English from L1 speakers was not how frequently they paused, but rather *where* they paused. The final important change in the measurement of fluency is that the quality of pauses produced are now often categorised as filled and unfilled (silent) pauses. SLA researchers (Clark & Fox Tree, 2002; Schmidt & Beers Fägersten, 2010) have suggested that while both silence and filled pauses indicate language processing demands, filled pauses also highlight emphasis, discourse organisation and communication strategies, helping facilitate communication (Dewaele, 1996; Schmidt & Beers Fägersten, 2010).

As for the repair aspect of fluency, it is possible to argue that the measurement of this aspect of fluency has changed the least. Repair measures have been historically calculated by counting the number of reformulations, false starts, self-corrections, repetitions, replacements or hesitations. Although this approach is still widely used to measure repair fluency, we identify at least two concerns with it. First, some of these measures overlap with one another or with other aspects of performance. False starts, for example, often lead to reformulations, and therefore the number of false starts and reformulations are internally dependent. Similarly, hesitations often precede or co-occur with other repair measures such as repetitions and replacements. Therefore, these measures may not provide a separate or independent representation of the repair phenomenon in fluency. Another problem is that the use of repair measures may be linked to personal speaking styles. For example, there is some initial evidence (e.g. Duran-Karaoz & Tavakoli, forthcoming) to suggest that L1 behaviour in repair measures is, at least to some extent, related to L2 repair strategies. Clearly, more research is needed to examine the nature of repair measures and the relationship between repair processes and other aspects of L2 performance.

An important decision to be made in any fluency study relates to which measures of fluency can best represent L2 utterance fluency. Our main concern is to investigate which measures most consistently and effectively demonstrate speech fluency in the context of the study and, therefore, it is crucial that measures are selected carefully. For example, studies have shown that some measures of fluency are inter-related and, if not chosen carefully, one measure may overlap with others (Kormos, 2006; Skehan, 2009, 2014; Tavakoli and Skehan, 2005). Measures should, therefore, be selected which hone in on specific aspects of fluency, and yet Skehan (2014) contends that in addition to such measures of speed, silence and repair, *composite* measures that blend *speed* and *flow* of speech, e.g. speech rate and length of run should also be considered.

Witton-Davies (2014) and Mora and Valls-Ferrer (2012) suggest that pause length, pause frequency, pause location, mean length of run, speech and articulation rates, phonation time ratio, and a selection of repair measures are the most reliable measures of utterance fluency. Prefontaine (2013) argues that mean length of run and average pause time are two measures of utterance fluency that most strongly relate to perceptions of fluency. Kahng (2014) suggests that speech rate and mean length of run are strongly associated with both L2 oral proficiency and perceived fluency, whereas articulation rate and repair measures are not. As discussed above, filled and unfilled pauses at mid-clause and end-clause positions are also potentially revealing measures of fluency and we agree with Hilton (2014) that composite measures of breakdown in which filled and unfilled pauses are combined may also help provide a better understanding of breakdown fluency. What emerges from the existing research, then, are a specific collection of measures which have been shown to reliably represent the construct of fluency as defined in Section 2. A complete list of measures used in this study and the related aspects of fluency that they represent are presented below in Section 7.3.

5. RESEARCH AIMS AND QUESTIONS

Building on recent understandings of L2 speech fluency in SLA, we aim to investigate how different aspects of fluency can most effectively be deployed in the Aptis operational rating scales. In effect, the current study aims to examine to what extent and in what ways test-takers output language validates the fluency descriptors used to define different levels of the Aptis Speaking test. Given the reported impact of task design on L2 performance, this research will aim to examine and identify fluency features (e.g. speed, pausing, or repair) that are most relevant to specific tasks and across different proficiency levels in the Aptis Speaking test.

The research questions guiding the study are:

RQ1: How are various aspects of fluency presented across different levels of proficiency (A2, B1, B2 and C1) in the Aptis Speaking test?

RQ2: To what extent is test-takers' fluency affected by task design (task type, discourse type and target level)?

6. ASSESSMENT OF SPEAKING IN APTIS AND RATING OF FLUENCY

The Aptis Speaking test consists of four parts targeting different CEFR levels, eliciting different types of output (see Table 2 below). The test structure with four parts, gradually increasing in difficulty, lends itself to a system of probing a test-takers' ability (O'Sullivan & Dunlea, 2015).

Part	Task	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.)

Table 2: Structure of the Aptis Speaking Test

Test-takers' performance on each task is examined by trained raters, who award a separate score to performance on each task. The Aptis rating system is innovative, as three different rating scales are used, according to the target proficiency level of each task (i.e. one scale for Part 1, one for Parts 2 and 3, one for Part 4), so that the rater can provide more accurate ratings for specific performance at each part.

The Aptis Speaking rating scale is holistic, but rating descriptors at each level include analytical points (e.g. fluency, vocabulary, grammar) to be weighted equally. As presented in Appendix 1, four different linguistic aspects of performance are featured in the three sets of the Aptis Speaking scales (*grammatical range and accuracy, lexical range and accuracy, pronunciation, fluency and cohesion*) in addition to *task fulfilment/topic relevance*. Performance descriptors are used to describe performance at each band. Fluency (combined with cohesion for the B1–B2 scales) is one of the main categories of linguistic ability assessed in this test (see Table 3 for a summary of assessed areas and example fluency descriptors).

Task	Target level	Areas assessed	Example fluency descriptors
1	A1/A2	**Areas assessed:** task fulfilment / topic relevance, grammatical range & accuracy, vocabulary range & accuracy, pronunciation, **fluency**	A1: Frequent pausing, false starts and reformulations impede understanding A2: Frequent pausing, false starts and reformulations but meaning is still clear
2	B1	task fulfilment / topic relevance, grammatical range & accuracy, vocabulary range & accuracy, pronunciation, **fluency** and cohesion	A1–2: Noticeable pausing, false starts and reformulations B1: Some pausing, false starts and reformulations
3	B1		
4	B2	task fulfilment / topic relevance, grammatical range & accuracy, vocabulary range & accuracy, pronunciation, **fluency** and cohesion	B2: Some pausing while searching for vocabulary but this does not put a strain on the listener C1: Backtracking and reformulations do not fully interrupt the flow of speech

Table 3: Fluency descriptors across tasks and proficiency levels in Aptis

The holistic approach to scoring is time-efficient, and represents a more natural, authentic way of judging people's speaking skills. However, it must be kept in mind that different raters may prioritise different aspects of the performance to arrive at their evaluation, thus potentially leading to less reliable results than analytic scoring (Taylor & Galaczi, 2011).

To address this concern, it is essential to ensure that the analytic descriptors included in each level of the holistic scales can be effectively applied to actual test-taker performance, as difficulties in matching descriptors and test-taker performance could be a potential threat to weighting all analytic points equally. This research, therefore, aims to provide performance benchmarks for fluency that can be used in the Aptis rater training as well as in refining rating descriptors (if necessary). Among various analytic criteria, the proposed focus on fluency in this study is believed to be most vital, since raters tend to show the least confidence in evaluating fluency (Brown, 2006b), and fluency seems to be the most susceptible to task elicitation methods (Nakatsuhara, 2012).

7. METHODOLOGY

As discussed above, over the past years, detailed linguistic and discourse analysis of test-taker language has become an informative method to examine whether test-takers' performance validates language descriptors used to define different levels of the rating scales (e.g. Brown et al., 2005). This approach, for example, has allowed researchers to investigate the extent to which these measures differentiate between adjacent levels of the rating scales. Building upon this methodology, the current study uses a range of measures to examine fluency in terms of speed, silence and repair dimensions of speech. Use of technical software, PRAAT (Boersma & Weenink, 2013), will ensure a more accurate measurement of the temporal aspects of fluency.

7.1. Research design

The study has a within and between-participant design of 2 x 2 with level of proficiency as a between-participant variable (4 levels of A2 to C1) and task type as a within-participant variable (Tasks 1-4). It should be noted that the A2 group rarely produced enough speech samples in Task 4 to carry out any meaningful analysis, and therefore the analysis of the results did not have any comparisons for this task for the A2 group. The factorial design allowed us to compare performances not only across different levels of proficiency and the four tasks, but to examine the possible interaction between task and level of proficiency.

7.2. Data set

The study examined 32 test-takers' audio-recorded performances across the four Aptis Speaking tasks, totalling 120 recordings (no A2 recordings on Task 4). Test-takers were selected so that there were 8 test-takers at each of the levels of proficiency (A2, B1, B2 and C1). With assistance from an experienced Aptis Speaking rater at the British Council, data were carefully selected in different bands from the test-takers whose overall, holistic scores represented their fluency scores across all 4 tasks. Jagged-profile test-takers across different components (e.g. Lexis, Grammar) of the holistic scales were avoided. The test-takers were both male and female and came from a range of different L1 background and nationalities, to minimise potential effects of test-taker characteristics on their speech features.

7.3. Analytic fluency measures

To carry out micro-analysis of fluency, fluency measures were carefully examined and selected. The range of the selected measures needed to be comprehensive as well as relevant to the candidates' output language designed to be elicited in the Aptis Speaking test. Based on examination of the current Aptis rating scales by the researchers and recommendations in SLA literature regarding various indices that can reliably measure different fluency aspects (e.g. De Jong et al., 2015; Kahng, 2015; Skehan, 2014), the following 20 analytic measures under three categories of speed, breakdown and repair were selected:

Speed measures

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

b) Articulation rate: total number of syllables divided by total amount of phonation time (excluding pauses) multiplied by 60

c) Mean length of run (pruned): the mean number of syllables between two pauses
(It should be noted that following de Jong et al., (2015) a pause is an unfilled silence of longer than 0.25 a second.)

Breakdown measures

a) Phonation time ratio: percentage of performance time spent speaking

b) Mean length of all silent pauses

c) Mean length of silent pauses at mid-clause *(f-1)* and end-clause *(f-2)* positions, respectively

d) Mean length of filled pauses at mid-clause *(g-1)* and end-clause *(g-2)* positions, respectively

e) Frequency of all silent pauses

f) Frequency of silent pauses at mid-clause *(j-1)* and end-clause *(j-2)* positions, respectively

g) Frequency of filled pauses

h) Frequency of filled pauses at mid-clause *(l-1)* and end-clause *(l-2)* positions, respectively

Repair measures

a) Frequency of all repairs (per 60 seconds)

b) Frequency of false starts and reformulations (per 60 seconds)

c) Frequency of partial or complete repetitions (per 60 seconds)

d) Frequency of self-corrections (per 60 seconds)

7.4. Data analysis procedures

The speech data were transcribed and then detailed linguistic and discourse analysis measures of fluency were used to examine the test-takers' speech. To achieve accurate measurement of fluency, the technical software 'PRAAT' (Boersma & Weenink, 2013) was used. The next section provides an operational description of how PRAAT operates, and a brief account of which of its functions were employed in this study.

7.5. Using PRAAT

PRAAT is often used in fluency research for its 'text grid to silences' feature (e.g. De Jong & Perfetti, 2011). This feature automatically detects silence in a speech sample. This was not possible in the current study because we were interested in both filled pauses, silence and pauses which combined silent and filled pause. However, PRAAT also allows for detailed manual investigation and annotation of speech samples and then automatic extraction of the duration of speech phenomena, such as pauses. An additional feature of PRAAT is that it allows researchers to write computer scripts for the kind of analysis they require. It is, therefore, a very precise, reliable and flexible tool for language research.

In the current study, all 120 task performances, totalling 4.2 hours of speech were converted to .WAV format which is compatible with PRAAT. One by one, these recordings were opened in PRAAT and were listened to at the same time as the spectrogram was studied. The spectrogram is accompanied by a 'text grid' which allows the researcher to annotate the speech sample. The analysis began at the first syllable uttered by the test-taker, be this of lexical content (e.g. 'My'), a filler (e.g. 'OK') or a non-verbal filler (e.g. 'um'). When identifying the beginnings and ends of runs of speech and pauses, the screen view was zoomed into at most .2 of a second resulting in very precise measurement.

Silent, filled and composite pauses were identified through repeated listening to small stretches of the recording accompanied by visual inspection of the spectrogram. Some silent pauses can be identified by inspection of the spectrogram, but often the picture is clouded by the test-taker's breathing (a sharp intake of breath can look like sound). Filled pauses can also display certain visual characteristics but these can easily be confused with syllable elongation which was not under investigation in the current study. Therefore, the spectrogram had to be studied in conjunction with careful listening to the speech. Only silences, non-verbal fillers or combinations of both which totalled .25 of a second or longer were marked as pauses. They were identified as either silent, filled or composite pauses on the text grid. The beginnings and ends of these pauses were marked against the spectrogram for the entire speech sample. This created alternating 'intervals' of speech and pause. Each pause interval was marked as either a filled, silent or composite pause. Each pause was then studied again, this time to ascertain the pause position (mid-clause versus end-clause). This was done by careful listening to the recordings and examination of orthographic transcription which had been marked with clause boundaries. Information about clause position was also added to the pause intervals on the text grid.

Between these pauses are the stretches or 'runs' of speech generated by the test-taker. These were listened to and studied visually in order to count the number of syllables produced. In most studies of L2 fluency, syllables are counted from orthographic transcriptions of the speech. It could be argued, however, that in spontaneous speech, especially that produced by language learners, syllables uttered do not conform with syllables expected. For example, in standard English 'student' is expected to have two syllables but a language learner may produce 'estudent' which totals three syllables. In order to avoid such complexities, manual counting of syllables using the original recording provides a more accurate (though by no means quick and simple) approach. Single runs may be listened to multiple times in order to ascertain number of syllables which were then added to the text grid. Any non-verbal filler shorter than .25 a second was counted as a syllable along with partially uttered words, repetitions, etc. Non-verbal phenomena, such as laughter, coughing and throat-clearing, was discounted from analysis altogether (i.e. it was not counted as a pause or part of a run). Any time spent laughing or coughing was also removed from sample time calculation so that it does not affect calculations of speed fluency. However, these phenomena often did mark the ends of runs. It is worth noting that such phenomena were incredibly rare in this data set, perhaps due to the monologic nature of the tasks.

The recordings in the APTIS oral examination are cut off at a set time (30 seconds for the questions in task 1; 45 seconds for each question in tasks 2 and 3; and 120 seconds for task 4). This meant that often a test-taker's speech was cut off in the middle of a run of speech. Where this happened, the analysis stopped at the previous run boundary, and any subsequent pause or interrupted run was removed from the analysis. For those test-takers who finished speaking before the allotted time was up, any time remaining after the last syllable was uttered was removed from the analysis.
It was unusual to remove more than a second or so for students at proficiency levels B1–C1. A2 level students, however, frequently abandoned the task after a few words or finished speaking before the time ran out and the recording stopped. For this reason, measures of speech rate and articulation rate for the A2 students are based on much smaller sample times. This is also expected to impact on other fluency measures, such as pause frequency and repair frequency. Results of the A2 level students, therefore, must be interpreted with care.

All recordings were analysed a second time. This time, repetitions, reformulations and self-corrections were identified and marked on the text grid. Here, length of repair was not important but these phenomena were added to the text grid in order to facilitate an automatic frequency calculation. 'False starts' and 'reformulations' were grouped together because of their interdependence.
All recordings were analysed a third time. This time pruned syllables were counted and marked on the text grid. Pruning involved discounting syllables which were any of the following:

- non-lexical fillers (um; er) shorter than 250ms
- syllables involved in repair (outlined above)
- lexical fillers (well; you know)
- epenthesis (e.g. the word studio pronounced estudio – in this example the 'e' would be pruned).

The 'pruned' syllable count was then added to the text grid below the 'unpruned' or 'raw' syllable count. When all the speech samples had been analysed in this way, a PRAAT script was written which would generate output that would allow us to calculate the various measures of fluency outlined. This script was run with each speech sample, generating 320 individual spreadsheets that provided information about the number and duration of all the phenomena that had been annotated. These were then merged into a single spreadsheet and the tool 'pivot table' was used to calculate the various fluency measures listed in Section 7.3 (see Glossary in Appendix 3 for descriptions).

8. ANALYSES AND RESULTS

Given the factorial design of the study, statistical analyses that allow for the examination of both within and between-participant effects were needed. A repeated measures MANOVA was, therefore, employed to explore the effects of task type, proficiency level, and the interaction between the two in the data set. Effect sizes were calculated to show the power of significant results. However, given the small sample size of the study, it was decided that using a Bonferroni adjusted alpha level, which would potentially increase Type 2 errors, was not appropriate. The descriptive statistics for all fluency measures across levels and tasks is provided in Appendix 2.

A repeated measures within-between participant MANOVA was run with 'Task' as the within-participant (four levels Tasks 1–4) and 'Level of proficiency' as the between-participant (four levels of A2, B1, B2 and C1) variable. The dependent variables included in this part of the analysis were chosen to demonstrate the different aspects of fluency, i.e. speed, breakdown and repair (see Section 7.3). Once the results of the repeated-measures MANOVA indicated significant differences, univariate analyses were used to examine the differences across different proficiency levels and tasks. We now present a summary of the results of the MANOVAs and report the major findings of the univariate analyses. For a simple description of the analytic measures, see the Glossary in Appendix 3.

8.1. Repeated-measures MANOVA

Checking multivariate normality through a linear regression, the results of Mahalanobis distances showed that our largest Mahal distance figure was 13.45, which is lower than the critical value of 26.13 suggested for an 8-dependent variable test. The results suggest there were no multivariate outliers in the dependent variables. (Note: Mahalanobis distance is the distance between a data point and a multivariate overall mean. It is a more powerful multivariate method for detecting outliers than examining one variable at a time because it considers the different scales between variables and the correlations between them.) Levene's Test of Equality of Error Variances showed that, for most measures, the assumption of equality of variance has not been violated. However, where this happened, we set a more conservative alpha level.

The Multivariate test showed three statistically significant differences for Level, Task and the interaction between the two independent variables. The significant differences were one for Proficiency Level (Wilks' Lambda = .160; F = 3.32, p = .000; η^2 = .457), one for Task (Wilks' Lambda = .280; F = 3.63, p = .008; η^2 =.720), and one for the interaction between Proficiency Level and Task (Wilks' Lambda = .097; F = 1.70, p = .04; η^2 =.540).

When the further results were considered, for the within-participant comparisons, three significant differences were observed. They were for repair measure (F = 14.31, p = .001; η^2 =.338), mean number of end-clause silent pauses (F = 5.77, p = .023; η^2 = .171), and mean length of mid-clause silent pauses (F = 6.24, p = .019; η^2 = .182). Test of between-participant comparisons (Proficiency Level) showed three significant differences for Proficiency Level: speech rate (F = 22.13, p < .001; η^2 = .703), mean length of mid-clause pauses (F = 16.99, p = .000; η^2 = .646), and mean length of end-clause pauses (F = 9.40, p = .000; η^2 = .502). These sets of results allowed us to continue with further analysis, e.g. univariates, to identify whether there were statistical differences between the various measures across tasks and proficiency levels.

8.2. Univariate analysis

Following from the repeated measures MANOVA results that showed significant differences in the data, a number of two-way between group analyses of variance were run to explore the effects of Level and Task type on different aspects of fluency. Tukey post-hoc comparison was used to identify the significant differences across the tasks and the levels. Given the purpose of the study, the research team decided to use parametric tests in all the analyses, despite the small size of samples in this research. In order to offer more informative results about the role of different aspects of fluency in candidates' speech, it was thought to be more important to avoid Type 2 errors rather than Type 1 errors. A summary of the major findings is presented below.

8.3. Fluency measures across proficiency levels and tasks

The results of the Univariate analyses are presented to show fluency measures across levels of proficiency and tasks. The findings are reported for the different aspects of fluency, i.e. speed, breakdown (silence) and repair measures respectively. It should be noted that since A2 candidates' performance on Task 4 was not included in the analysis, the analysis is drawing on a lower number of performances on Task 4, which may have had an impact on the results obtained.

8.3.1. Speed measures

In this section, the results of the analyses will be presented for measures of a) speech rate, b) articulation rate, c) mean length of run and d) phonation time ratio. Each analysis is shown by a corresponding figure.

a) Speech rate:

A significant difference was observed for speech rate across different levels of proficiency ($F = 59.19$, $p < .001$, $\eta^2 = .628$). The post hoc analysis showed that A2 level was different from B1, B2 and C1. B1 level was also different from B2 and C1. However, B2 and C1 levels were not statistically different. The order of the speed with which the test-takers at different levels spoke was C1 > B2 > B1 > A2. The analysis did not show a significant difference across the tasks, suggesting the speed with which the four tasks were performed was comparable across the four tasks. The order of speed across tasks was T4 > T3 > T2 > T1.

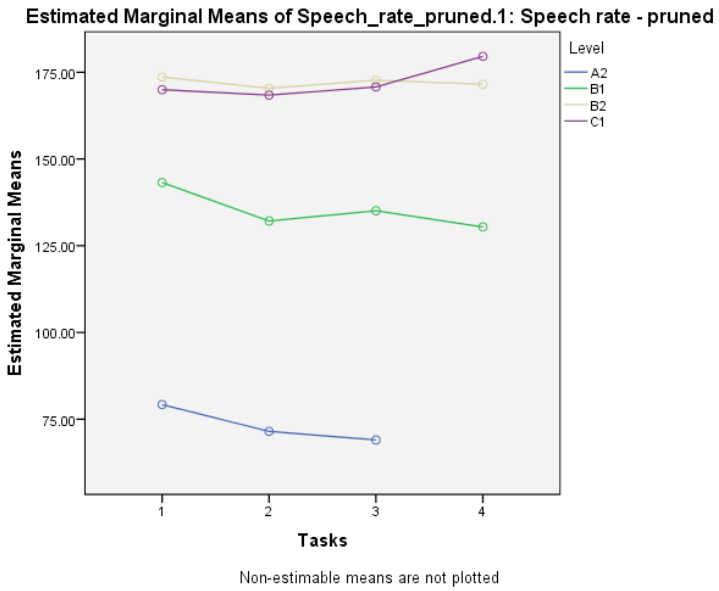

Figure 1: Speech rate across levels and tasks

b) Articulation rate:

A significant difference was observed for articulation rate across different levels of proficiency ($F = 30.63$, $p < .001$, $\eta^2 = .467$). The post hoc analysis showed that A2 and B1 levels were different from each other and from B2 and C1. However, B2 and C1 levels were not statistically different. The order of speed with which the test-takers performed the speaking tasks was C1 > B2 > B1 > A2. The results did not show a significant effect of task, suggesting speed of performance across different tasks was consistent.

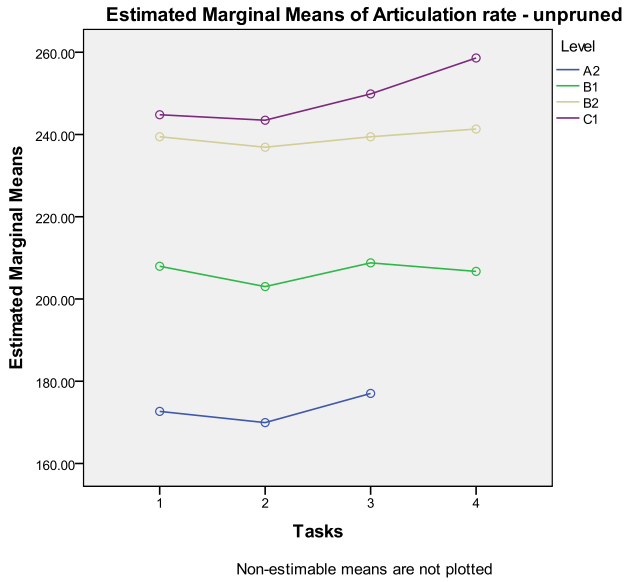

Figure 2: Articulation rate across levels and tasks

c) Mean length of run:

A significant difference was observed for mean length of run across different levels of proficiency ($F = 46.51$, $p < .001$, $\eta^2 = .571$). The post hoc analysis showed that A2 and B1 levels were different from each other and from B2 and C1. However, B2 and C1 levels were not statistically different. The order of the length of run in the test-takers' speech was B2 > C1 > B1 > A2. As for the effects of tasks, the results suggested there was not a statistically meaningful difference between the mean length of run across different tasks. The order was T4 > T3 > T2 > T1.

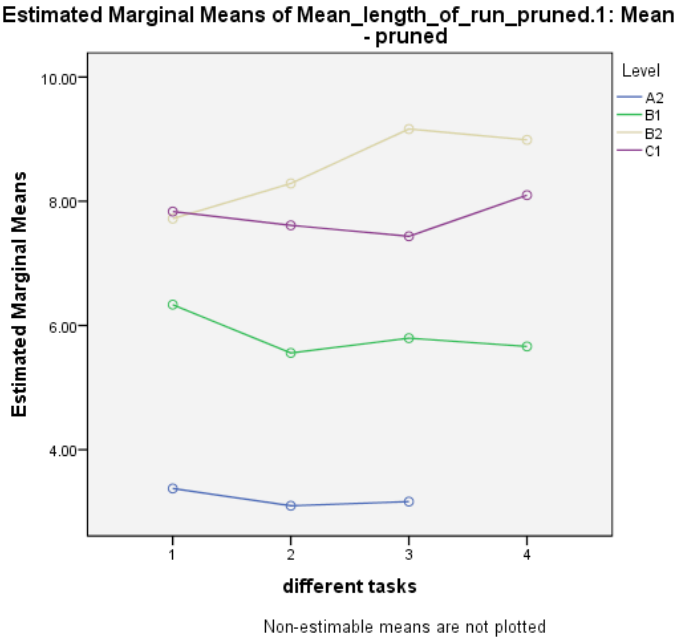

Figure 3: Mean length of run across levels and tasks

d) Phonation time ratio:

A significant difference was observed for phonation time ratio across different levels of proficiency ($F = 85.47$, $p < .001$, $\eta^2 = .710$). 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. The order of phonation time ratio among groups was B2 > C1 > B1 > A2. The analysis did not show a significant result across the tasks, suggesting the phonation time ratios were comparable across the four tasks. The order of phonation time ratios across the tasks was T4 > T1 > T2 = T3.

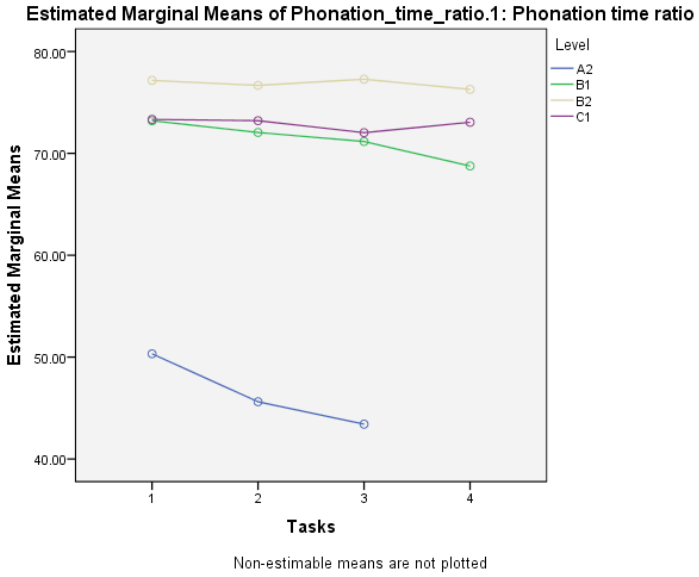

Figure 4: Phonation time ratio across levels and tasks

The results of the univariate analysis showed a significant difference in the total amount of time spoken across the different tasks. We do not consider this an important task effect to be reported here as different time allocations are considered for the different tasks in the Aptis Speaking test (see Section 6).

8.3.2. Breakdown measures

As discussed above, we have used a relatively large number of breakdown measures to capture the full picture of how the breakdown phenomenon affects fluency across different proficiency levels and tasks. Silent, filled and composite pauses are examined in terms of their length and frequency; and pauses are examined with regard to their location, i.e. whether they are located in mid-clause or end-clause position. Below, we will first present measures that focus on length of pauses. We will then look at the measures that examine the frequency of pauses.

Length of pauses

e) Total length of all pauses:

A significant difference was observed for total length of pauses across different proficiency levels ($F = 50.28$, $p < .001$, $\eta^2 = .590$). The post hoc analysis showed that the A2 level was different from B1, B2 and C1. B1, B2 and C1 were not different from one another. The order of length of pauses at different levels was A2 > B1 > C1 > B2. There were no significant differences in the total length of pauses across different tasks.

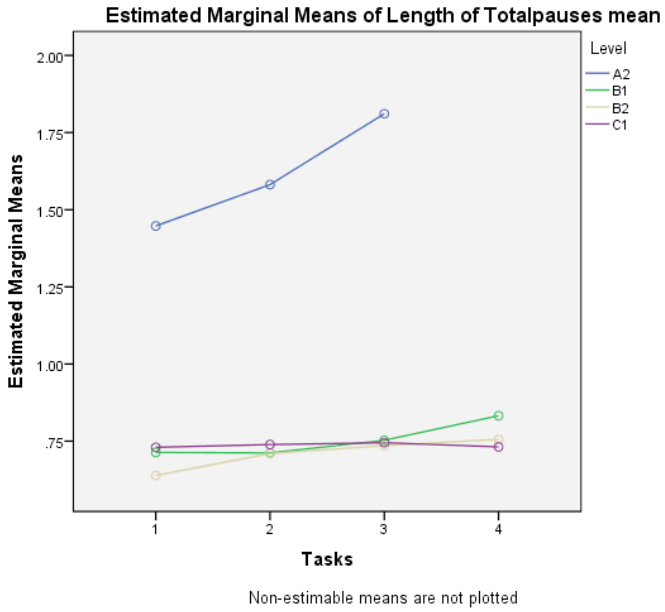

Figure 5: Total length of pauses across levels and tasks

f-1) Mean length of mid-clause silent pauses:

A significant difference was observed for length of mid-clause silent pauses across different proficiency levels ($F = 30.47$, $p < .001$, $\eta^2 = .465$). The post hoc analysis showed that the A2 level was different from B1, B2 and C1. The other levels were not statistically different from one another. The order of length of mid-clause silent pauses at different levels was A2 > B1 > B2 > C1. For the differences across the four tasks, no significant differences were observed.

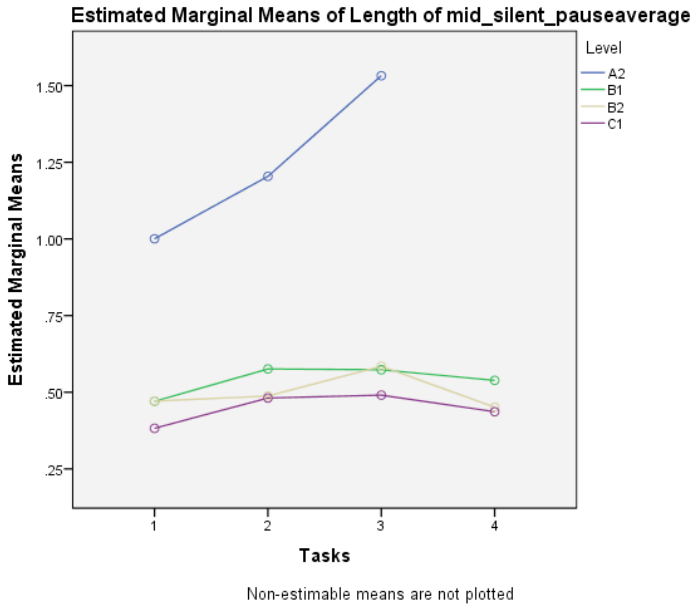

Figure 6: Length of mid-clause silent pauses across levels and tasks

f-2) Mean length of end-clause silent pauses:

A significant difference was observed for length of end-clause silent pauses across different levels of proficiency ($F = 26.37$, $p < .001$, $\eta^2 = .430$). The post-hoc analysis showed that A2 level was different from B1, B2 and C1. The other levels were not different from one another. The order of length of end-clause silent pauses at different levels was A2 > B1 > B2 > C1. No significant differences were observed for the length of end-clause silent pauses across the four tasks.

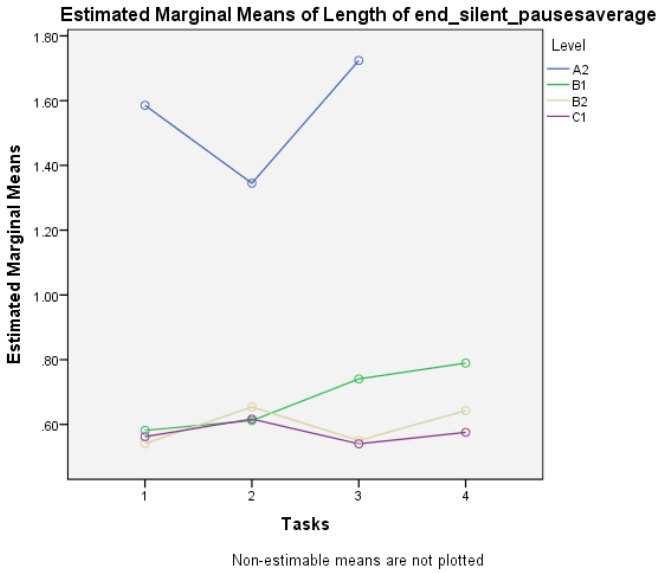

Figure 7: Length of end-clause silent pauses across levels and tasks

g-1) Mean length of mid-clause filled pauses:

No significant difference was observed for mean length of mid-clause filled pauses across different levels of proficiency. Interestingly, the length of mid-clause filled pauses was longer at higher proficiency levels, suggesting more proficient speakers use longer filled pauses. In particular, C1 and B1 speakers appeared to use longer filled pauses in Tasks 3 and 4. The order for length of mid-clause filled pauses at different levels is C1 > B1 > B2 > A2. No significant differences were observed across tasks.

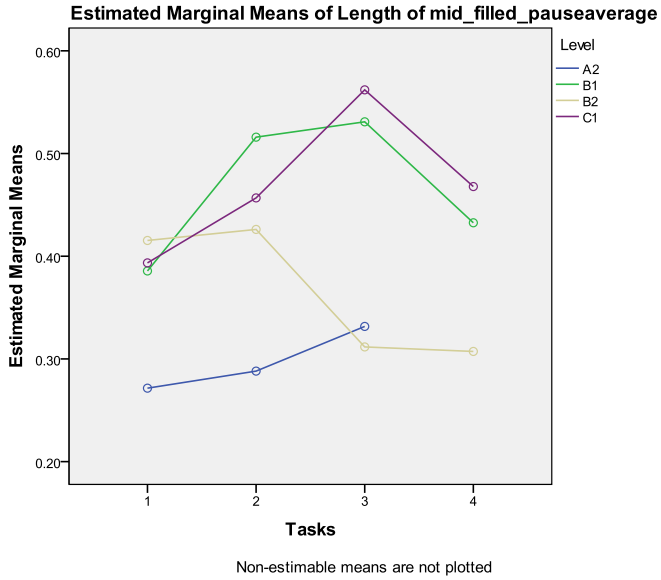

Figure 8: Length of mid-clause filled pauses across levels and tasks

g-2) Mean length of end-clause filled pauses:

No significant difference was observed across different levels of proficiency. Similar to the results for length of mid-clause filled pauses, the length of end-clause filled pauses was longer at higher proficiency levels. The length of end-clause filled pauses at different levels, mirrored the mid-clause filled pauses pattern, i.e. the order was C1 > B1 > B2 > A2. No significant results were observed across the four tasks.

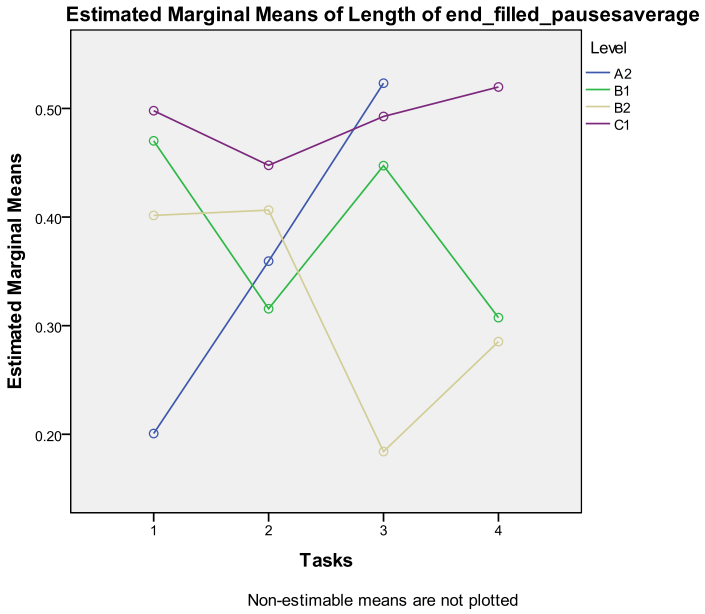

Figure 9: Length of end-clause filled pauses across levels and tasks

Frequency of pauses

i) Total number of silent pauses:

Although the results of the univariate analysis did not show a significant effect of level of proficiency on number of silent pauses ($p < .186$), the figures clearly suggest that there are more silent pauses in lower levels of proficiency (A2 = 29.11, B1 = 28.63, B2 = 24.70, and C1 = 23.64). The results were not significant with regard to the effect of task type ($p < .858$).

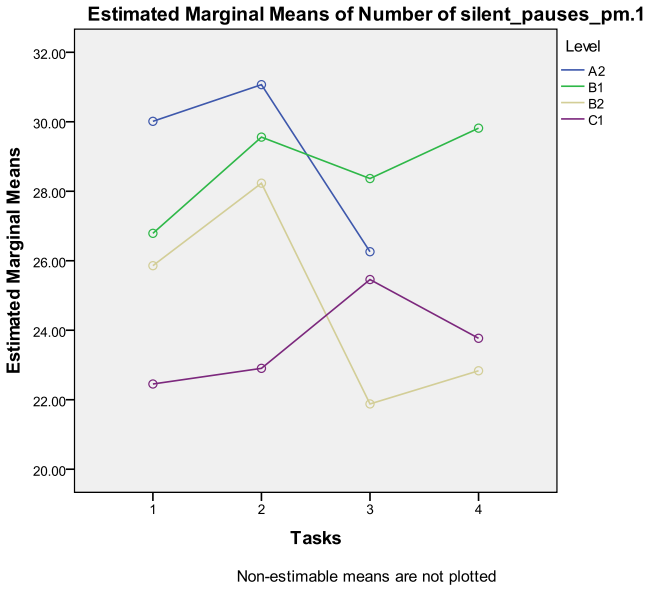

Figure 10: Number of silent pauses across levels and tasks

j-1) Mean number of mid-clause silent pauses:

A significant difference was observed for number of mid-clause silent pauses across different levels of proficiency ($F = 7.17$, $p < .001$, $\eta^2 = .170$). 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. The order for number of mid-clause silent pauses at different levels was A2 > B1 > C1 > B2. Once again, the B2 level is not following the progressive pattern as they pause less frequently than C1 level. The results suggested that the number of mid-clause silent pauses was not different across the tasks.

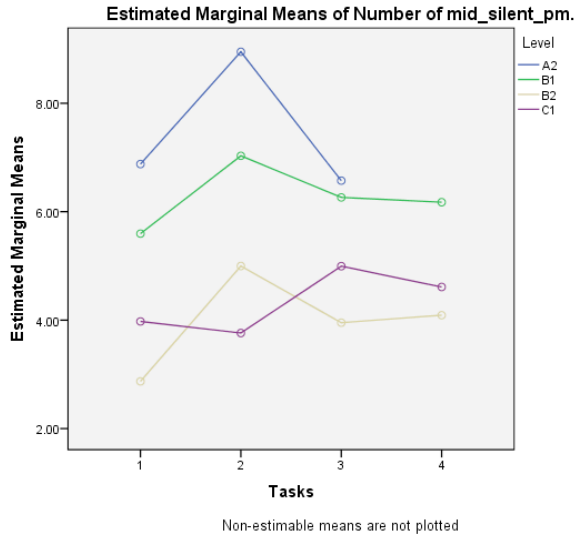

Figure 11: Number of mid-clause silent pauses across levels and tasks

j-2) Mean number of end-clause silent pauses:

No significant difference was observed for number of end-clause silent pauses across different levels of proficiency ($p < .535$). The number of end-clause silent pauses at different levels was very similar (A2 = 7.08; B1 = 8.04; B2 = 8.37; and C1 = 7.48). No significant results were observed for differences across the tasks ($p = .723$).

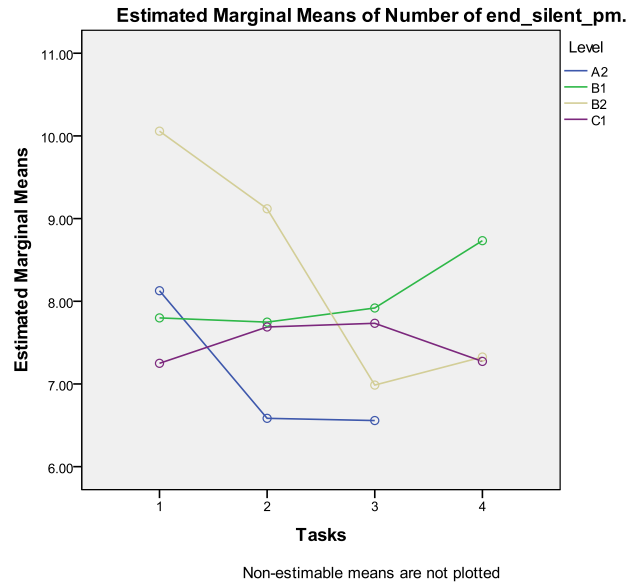

Figure 12: Number of end-clause silent pauses across levels and tasks

k) Total number of filled pauses:

A significant difference was observed for total number of filled pauses across different levels of proficiency ($F = 4.04$, $p < .009$, $\eta^2 = .103$). The post hoc analysis showed that A2 was different from B1 and C1, but not from B2. B1, B2 and C1 were not different from each other. The order for number of filled pauses at different levels was B1 > C1 > B2 > A2. This result suggests that, with the exception of the B1 level, test-takers at higher levels of proficiency produced more filled pauses. No significant results were obtained for differences across the tasks ($p < .682$).

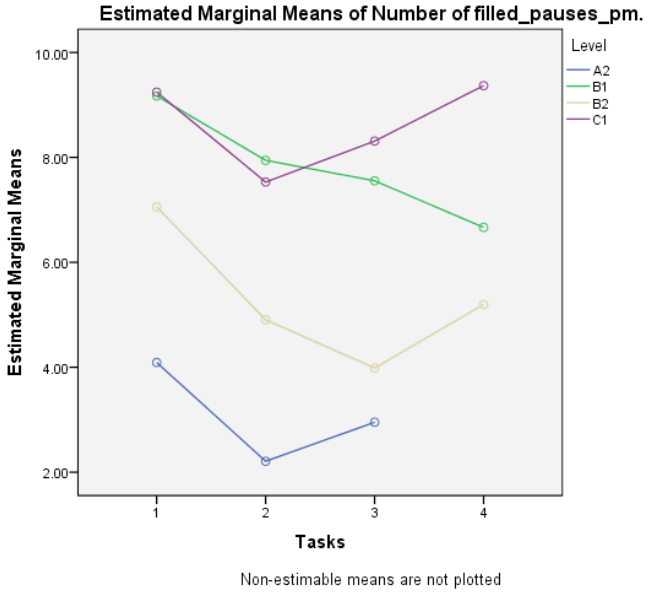

Figure 13: Total number of filled pauses across levels and tasks

l-1) Mean number of mid-clause filled pauses:

A significant difference was observed for number of mid-clause filled pauses across different levels of proficiency ($F = 4.76$, $p < .004$, $\eta^2 = .120$). The post hoc analysis showed that A2 was different from B1 and C1, but not from B2. B1, B2 and C1 were not different from each other. The order for number of filled pauses at different levels was C1 > B1 > B2 > A2. This result suggests test-takers at higher levels of proficiency produced more mid-clause filled pauses. No significant results were obtained for differences across the tasks ($p < .667$).

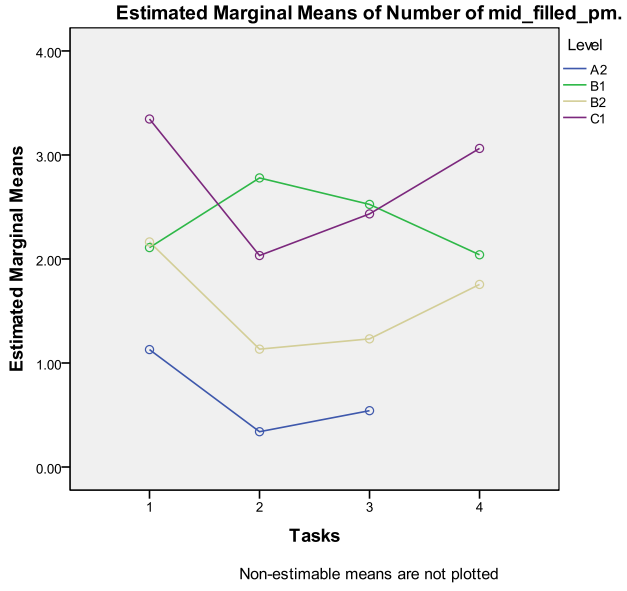

Figure 14: Number of mid-clause filled pauses across levels and tasks

I-2) Mean number of end-clause filled pauses:

The results of the univariate analysis did not show any significant differences across proficiency levels ($p < .151$) or the four tasks ($p < .729$). The emerging pattern suggests that test-takers at a higher proficiency level tend to use more filled pauses (A2 = .87; B1 = 1.55, B2 = 1.10; and C1 = 1.58).

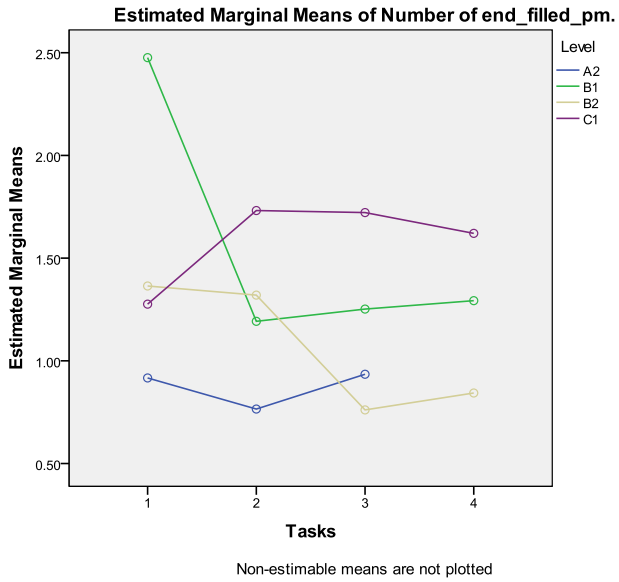

Figure 15: Number of end-clause filled pauses

8.3.3. Repair measures

m) Total number of repair measures:

A significant difference was observed for total number of repair measures across different proficiency levels ($F = 4.34$, $p < .006$, $\eta^2 = .110$). 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. The order for total number of repair measures is B1 > C1 > B2 > A2 (A2 = 4.04; B1 = 9.25; B2 = 7.63; C1 = 8.06).

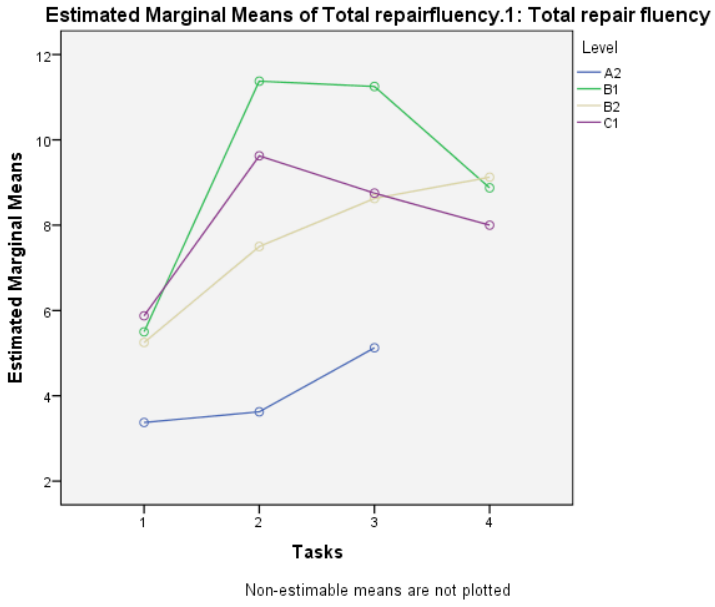

Figure 16: Total number of repair measures across levels and tasks

The results also indicated a significant difference across tasks ($F = 2.80$, $p < .04$, $\eta^2 = .07$). The post-hoc analysis indicated that Tasks 1 and 3 were different from one another. The results showed that repairs elicited by Tasks 2, 3 and 4 were very similar, but those elicited by Task 1 were much lower in numbers. The order was T4 > T3 > T2 > T1.

n) Mean number of false starts and reformulations:

A significant difference was observed for reformulations across different levels of proficiency ($F = 6.56$, $p < .001$, $\eta^2 = .158$). The post-hoc analysis showed that A2 level was different from B1 but not from other levels, suggesting speakers at the two ends of the proficiency continuum may be less active in reformulating their utterances. There was no significant difference between B1, B2 and C1 levels. The order of the use of reformulations at different levels is B1 > B2 > C1 > A2. No significant differences were observed in the reformulations across different Tasks ($p < .102$).

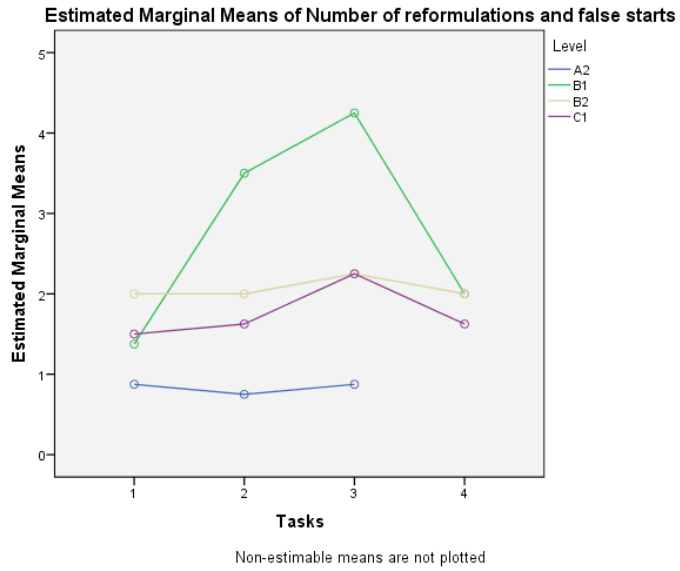

Figure 17: Number of false starts and reformulations across levels and tasks

o) Mean number of repetitions:

No significant difference was observed for repetitions across different levels of proficiency ($p < .236$) or tasks ($p < .077$). However, an interesting pattern emerged here. The number of repetitions increased from Task 1 to Tasks 2, 3 and 4 for A2, B1 and B2 levels. The C1 level produced more repetitions in Task 2 and fewer in Tasks 3 and 4. The order with which repetitions occurred were T4 > T3 > T2 > T1. In general, it can be argued that higher proficiency levels produced more repetitions (C1 > B1 > B2 > A2). Task 4 was the most and Task 3 the least repetitive task.

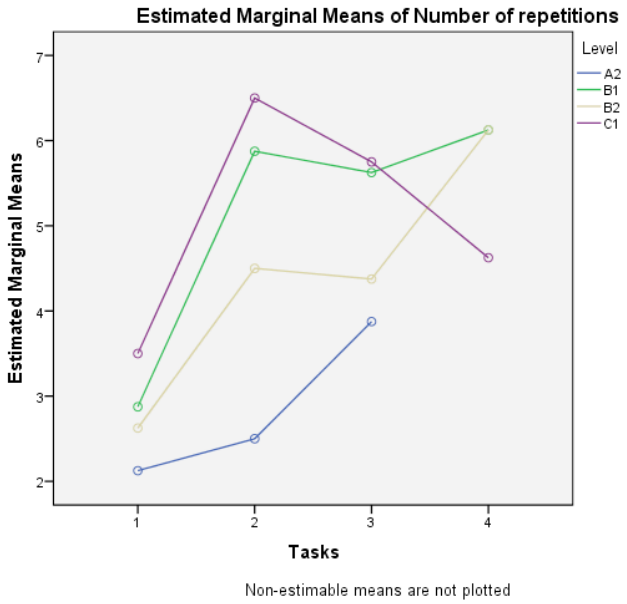

Figure 18: Number of repetitions across levels of proficiency and tasks

p) Mean number of self-corrections:

No significance difference was observed for the number of self-corrections across proficiency levels ($p < .06$) or tasks ($p < .61$). Similar to the results for other repair measures, B1 level produced the most and A2 level the fewest self-corrections. The order of self-corrections in different levels was B1 > C1 > B2 > A2. While a clear pattern was not observed across tasks, it was interesting to see A2 level consistently made very few self-corrections across different tasks.

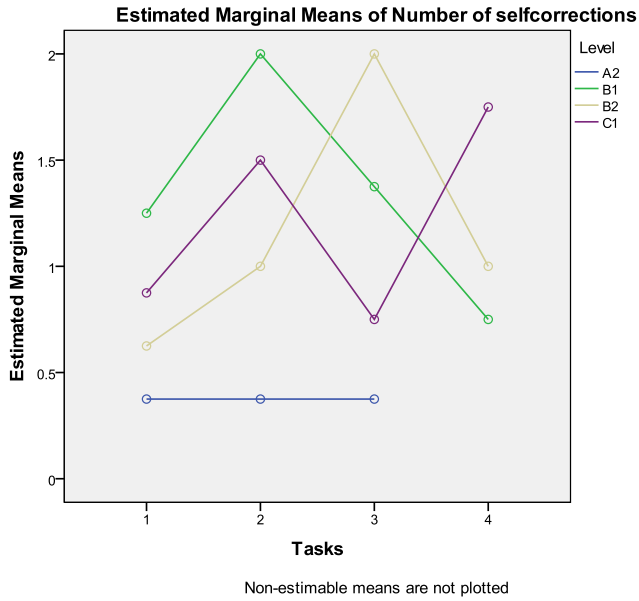

Figure 19: Mean number of self-corrections

9. DISCUSSION OF THE FINDINGS

In this section, we will present a summary of the results to show the extent to which aspects of fluency, i.e. speed, breakdown and repair, are affected by levels of proficiency and different task types. Each summary is then followed by a discussion of the findings.

Speed measures

Overall, measures of speed fluency distinguish performance at different levels of proficiency. All measures of speed distinguish A2 level from other levels. Measures of articulation rate, speech rate and mean length of run also distinguish fluency between B1 on the one hand and B2 and C1 on the other. All the results suggest that speed fluency is not statistically different between B2 and C1 level. The effect sizes for these comparisons, ranging from .457 to .710, are considered medium-size effects (Cohen, 1989), implying a noticeable degree of the variance in the speed of performances was related to the effect of proficiency level.

The results suggest that two levels of B2 and C1 are not statistically different in terms of speed fluency, although B2 level shows more speed in mean length of run and phonation time ratio, and C1 level shows faster articulation rate. The lack of distinction between the speed of 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. Alternatively, it is possible to argue that the results imply that a more demanding task at the C1 level may be needed in order to distinguish the speed performance of B2 and C1 level candidates. It is important to note that while articulation rate excludes pauses and, therefore, provides a more speed-only view of fluency, speech rate and mean length of run are speed measures that combine pausing and speed and, as such, they provide a more complete profile of the speaker fluency (Skehan, 2014).

As for the effects of task type, speed fluency does not seem to be affected by task type at a statistically meaningful level. The emerging patterns suggest that Task 4 is usually produced with more speed. However, this increase in the speed can be explained by the nature of the task (i.e. an extended piece of monologue) or in terms of practice effect (i.e. the candidate has already completed three other tasks and may be more prepared emotionally and psychologically). This can only be confirmed if data are collected through a counter-balanced design. It is also possible to argue that the tasks are too similar, in their cognitive demands, to reflect differences in speed fluency.

Breakdown measures: Length of pauses

The analysis of length of silent pauses examined silent and filled pauses. The results of the analysis suggested that length of silent pause is a measure that consistently distinguishes A2 level from other levels of proficiency, while B1, B2 and C1 levels are not different from each other as regards length of silent pauses. For all measures of length of silent pauses, A2 level produced the longest pauses (total pause, mid-clause and end-clause). The effects sizes for these comparisons ranged from .43 to .52. Interestingly, the differences between length of silent pauses show a progressive pattern of increase from A2 to C1, suggesting that length of silent pauses can potentially reflect proficiency levels.

With length of filled pauses, although none of the comparisons demonstrated a significant difference across different levels, a meaningful pattern emerged where speakers at higher proficiency levels produced longer filled pauses. The results show that test-takers at C1 level had the longest and at A2 level the shortest filled pauses. These results are in line with Revesz et al. (2016) who reported filled pauses could predict proficiency level.

Breakdown measures: Frequency of pauses

The results of the analysis of number of pauses reveal a number of interesting findings. First, unlike the findings for length of silent pauses, analyses of the *number* of silent pauses do not present a consistent pattern. While number of mid-clause pauses indicated a significant difference, total number of pauses and number of end-clause pauses did not. Number of mid-clause silent pauses distinguished lower levels of proficiency (A2 and B1) from higher levels (B2 and C1), suggesting that lower level candidates are silent more frequently at mid-clause positions. Overall, number of mid-clause silent pauses decreased with an increase in proficiency. The number of end-clause silent pauses did not distinguish among proficiency levels, implying that test-takers' number of pauses at end-clause positions was similar across different proficiency levels. This finding is in line with previous research findings that claim frequency of *mid*-clause pausing is a characteristic of L2 speech. However, given the limited evidence provided here, we suggest it is interpreted with caution.

Regarding the number of *filled* pauses, results revealed 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 candidates did not fit the same pattern, often using as many filled pauses as C1 candidates. 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. The significant results for number of mid-clause filled pauses indicated that candidates at higher proficiency levels use more filled pauses in general, and more mid-clause filled pauses in particular. In fact, it is interesting to see that the more proficient speakers use mid-clause filled pauses more frequently, whereas less proficient speakers produce more mid-clause *silent* pauses. Once again, no significant differences were obtained with regard to number of pauses across different tasks. Neither was a clear pattern observed. 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 notable contribution to the understanding of breakdown fluency in our field.

Repair measures

The data analysis indicated statistically significant differences across proficiency levels in their use of reformulations and total number of repairs. The results showed that B1 level produced the most and A2 level the least number of total repairs, repetitions and reformulations. Given the very few repair measures observed at A2 level, one way to interpret this finding is to argue that for L2 speakers to engage with repair processes, having achieved a certain proficiency level, i.e. a B1 level, may be necessary. As proficiency increases to B1 and B2 level, more repair processes are activated, and when they reach C1 level, the candidates use repair measures more reasonably and in moderation. 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.

Regarding the effects of task type on repair fluency, the results indicated a significant difference between Task 1 and other tasks in terms of total repair measures, but the effect size is small (.07). Overall, in comparison Task 1 elicits the fewest and Task 4 the most repair measures.

Table 4 summarises the results of task and level comparisons. In this summary 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. Some remarks are also included in square brackets.

Speed measures	**Level**	**Task**
a) Speech rate	(C1 = B2) > B1 > A2	No diff.
b) Articulation rate	(C1 = B2) > B1 > A2	No diff.
c) Mean length of run	(B2 = C1) > B1 > A2	No diff.
Breakdown measures	**Level**	**Task**
d) Phonation time ratio	(B2 = C1 = B1) > A2	Not diff.
e) Mean length of all silent pauses	A2 > (B1 = B2 = C1)	No diff.
f-1) Mean length of mid-clause silent pauses	A2 > (B1 = B2 = C1)	No diff.
f-2) Mean length of end-clause silent pauses	A2 > (B1 = B2 = C1)	No diff.
g-1) Mean length of mid-clause filled pauses	No diff. [longer filled pauses by C1 & B1 esp. in Tasks 3 & 4]	No diff.
g-2) Mean length of end-clause filled pauses	No diff. [longer filled pauses by C1 esp. in Task 4]	No diff.
i) Frequency of silent pauses	No diff. [more silent pauses at lower levels]	No diff.
j-1) Frequency of mid-clause silent pauses	(A2 = B1) > (C1 = B2)	No diff.
j-2) Frequency of end-clause silent pauses	No diff.	No diff.
k) Frequency of filled pauses	B1 = C1 = B2, C1 > A2, B1 > A2, B2 = A2	No diff.
l-1) Frequency of mid-clause filled pauses	C1 = B1 = B2, C1 > A2, B1 > A2, B2 = A2	No diff.
l-2) Frequency of end-clause filled pauses	No diff. [more filled pauses by C1 and B1 esp. in Tasks 3 & 4]	No diff.
Repair measures	**Level**	**Task**
m) Frequency of all repairs	(B1 = C1) > A2, B2 = B1 = C1,	T4 = T3 = T2, T3 > T1
n) Frequency of false starts & reformulations	(B1 = B2 = C1) > A2	No diff.
o) Frequency of repetitions	No diff. [more repetitions by C1 & B1 in Task3; by B1 & B2 in Task 4]	No diff. [more repetitions by A2, B1 & B2 as the task proceeds, but fewer repetition by C1 in Task 4]
p) Frequency of self-corrections	No diff. [very few at A2 level and many at B1 level]	No diff.

Table 4: Summary of the level and task comparisons of all analytic measures

In light of the discussion of the findings above, the two research questions of this study are now addressed.

RQ1: How are various aspects of fluency presented across different levels of proficiency (A2, B1, B2, and C1) in the Aptis Speaking test?

A summary of the most important findings discussed above suggests the following findings.

- Speed fluency distinguishes A2, B1 and B2 levels reasonably consistently. B2 and C1 levels are usually not different in terms of speed fluency.
- Length of silent pauses distinguishes A2 level from other proficiency levels.
- Length of mid-clause filled pauses, although not at a significant level, is longer in higher proficiency levels (except the B1 group, who produces longer filled pauses than B2).
- Frequency of mid-clause silent pauses distinguishes lower (A2 and B1) from higher (B2 and C1) proficiency levels.

- Frequency of filled pauses (both total number and mid-clause pauses) distinguishes A2 from higher levels.
- Higher proficiency levels use filled pauses more frequently than lower levels.
- Repair measures (both total number and 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.

These findings were encouraging as the Aptis Speaking test can utilise the above fluency characteristics as criterial features of each band level, in order to validate or modify the fluency rating descriptors of the test. However, a concern was raised in relation to the difficulty in differentiating B2 and C1 candidates in terms of their fluency performance. While the results indicated some straightforward fluency characteristics that can differentiate A2 from B1, B1 from B2, the results failed to identify a useful measure to distinguish B2 and C1 performances. As noted earlier, one possible way to interpret this is a ceiling effect which comes into play at the B2 level for many of the fluency aspects. This would mean that what makes C1 candidates different from B2 candidates may be, for example, the use of more sophisticated vocabulary and complex grammatical structures rather than how 'fluent' they are. Another interpretation is that the Aptis Speaking test which has a B2 task (Task 4) but lacks a C1 task is not capable of pushing B2 and C1 candidates to their linguistic limit for fluency. The lack of a more demanding task at C1 might, therefore, be preventing the test from capturing differential fluency performances that could be elicited from B2 and C1 candidates.

The second research question of our study was as follows.

RQ2: To what extent is test-takers' fluency affected by task design (task type, discourse type and target level)?

A summary of the above discussion suggests the following findings.

- Speed of performance is not affected by task type.
- Length of pauses is not affected by task type.
- Frequency of pauses is not affected by task type.
- Repair measures distinguish Task 3 from Task 1. Task 3 elicits most repairs.

These results imply that the performance is largely not affected by task type. Given the literature on the impact of task design on elicited fluency features (e.g. Michel, 2011; Robinson, 2007; Tavakoli, 2016, see Section 2 for more details), this finding was rather surprising and counter-intuitive. However, this may imply that the four Aptis tasks are not distinctive enough to impose different types of demand on the candidates' cognitive processes to affect their fluency performance. Table 5 presents Field's (2011) cognitive processing model of speaking based on Levelt (1989).

Information sources feeding into phases of the processing system	Cognitive processes	Outputs of processing
Speaker's general goals World knowledge Knowledge of listener Knowledge of situation Recall of discourse so far Rhetoric and discourse patterns	**CONCEPTUALISATION**	Pre-verbal message
Recall of ongoing topic Syntax Lexical knowledge Pragmatic knowledge Knowledge of formulaic chunks Combinatorial possibilities (syntactic/ collocational)	**GRAMMATICAL ENCODING:** *constructing a syntactic frame forming links to lexical entries*	Abstract surface structure
Lexical knowledge Phonological knowledge	**MORPHO-PHONOLOGICAL ENCODING:** *conversion to linguistic form*	Phonological plan
Syllabary: Knowledge of articulatory settings	**PHONETIC ENCODING:** *conversion to instructions to articulators; cues stored in a speech buffer.*	Phonetic plan
	ARTICULATION: *execution of instructions*	Overt speech
Speaker's general goals Target utterance stored in buffer Recall of discourse so far	**SELF-MONITORING**	Self-repair

Table 5: Cognitive processing model of speaking ability (Field, 2011: 74–77)

All the four tasks in the Aptis Speaking test are monologic tasks to respond to computer-delivered prompts. These prompts in Task 1 include only aural and written input while the remaining tasks also have non-verbal input, and Task 4 includes a pre-task planning time. These factors are likely to affect the candidates' *conceptualisation* process (the first stage in Table 5). Cognitive demands in terms of *grammatical encoding* (the second stage) in the four Aptis tasks seem to be graded by the language functions targeted in each task (O'Sullivan and Dunlea, 2015: 52–55). As such, tasks are indeed graded to have different levels of cognitive demand. However, these differences in tasks do not seem to be as distinctive as those used in the previous fluency studies (e.g. monologic and dialogic tasks in Tavakoli, 2016).

The lack of tasks distinctive enough to elicit differential fluency features across different parts of the test, however, is not a negative finding. This does not invalidate the Aptis Speaking test or its by-part rating system. This simply indicates that the three different scales in the Aptis Speaking test and the by-part rating system are useful, not because the tasks elicit different types of fluency performance, but because they elicit different levels of fluency performance, making it easier for the examiners to focus on narrower boundaries in making judgements. The use of the common scale between Task 2 and Task 3, both of which target the B1 level, is therefore justified.

10. RECOMMENDATIONS

Following the discussion presented in the previous section, we propose some modifications in the Aptis speaking rating scales and rater training materials. The recommendations outlined under Sections 10.1 and 10.2 below were presented to the British Council's Assessment Research Group on 26 May 2017. The meeting was held to discuss the recommendations derived from the findings of this research, in order to explore how best our findings could inform possible revisions of the Aptis Speaking test. To ensure the practical value of this study, it was thought to be significant to gather the test provider's voices at this final stage of the project. Therefore, the following recommendations reflect some modifications suggested and agreed in the meeting.

10.1 Recommendations for the Aptis Speaking rating scales

The following tables (Tables 6, 7 and 8) illustrate both the current and recommended fluency descriptors for Task 1, Tasks 2 and 3, and Task 4. The criterial features found to be useful to differentiate each level are incorporated in the modified descriptors below. In doing so, efforts were made to adhere to the five criteria for effective descriptor formulation proposed by the Council of Europe (2001: 205–207). The five criteria are:

- **Positive:** positively worded, describing what the test-taker can do
- **Definite:** The examiner is able to confirm clearly that 'yes' the test-taker has shown evidence that he or she can do what is described, or 'no' he or she cannot
- **Clear:** jargon-free and readily interpretable by assessors
- **Brief:** Any descriptor longer than a two-clause sentence cannot be used in the course of an operational assessment
- **Independent:** The assessor should not need to refer to other points on the scale in arriving at a decision concerning whether or not a test performance matches a descriptor

5 **B1 (or above)**	**Current**	Likely to be above A2 level.
4 **A2.2**	**Current**	Frequent pausing, false starts and reformulations but meaning is still clear.
	Modified	**Slow speed of speech and long silent pauses but meaning is still clear.**
3 **A2.1**	**Current**	Frequent pausing, false starts and reformulations but meaning is still clear.
	Modified	**Slow speed of speech and long silent pauses but meaning is still clear.**
2 **A1.2**	**Current**	Frequent pausing, false starts and reformulations impede understanding.
	Modified	**Slow speed of speech and long silent pauses impede understanding.**
1 **A1.1**	**Current**	Frequent pausing, false starts and reformulations impede understanding.
	Modified	**Slow speed of speech and long silent pauses impede understanding.**
0 **A0**	**Current**	No meaningful language or all responses are completely off-topic (e.g. memorised script, guessing).

Table 6: Suggested fluency descriptors for Task 1

5 **B2 (or above)**	**Current**	Likely to be above B1 level.
4 **B1.2**	**Current**	Some pausing, false starts and reformulations.
	Modified	**Moderate speed of speech but interrupted by mid-clause pauses and reformulations.**
3 **B1.1**	**Current**	Some pausing, false starts and reformulations.
	Modified	**Moderate speed of speech but interrupted by mid-clause pauses and reformulations.**
2 **A2.2**	**Current**	Noticeable pausing, false starts and reformulations.
	Modified	**Slow speed of speech and long silent pauses.**
1 **A2.1**	**Current**	Noticeable pausing, false starts and reformulations.
	Modified	**Slow speed of speech and long silent pauses.**
0	**Current**	Performance below A2, or no meaningful language or the responses are completely off-topic (e.g. memorised script, guessing).

Table 7: Suggested fluency descriptors for Tasks 2 and 3

5 **C1**	**Current**	Backtracking and reformulations do not fully interrupt the flow of speech.
	Modified	**Natural speed of speech, with some filled pauses and reformulations used effectively.**
4 **B2.2**	**Current**	Some pausing while searching for vocabulary but this does not put a strain on the listener.
	Modified	**Natural speed of speech, with some pauses and reformulations that do not interrupt the flow.**
3 **B2.1**	**Current**	Some pausing while searching for vocabulary but this does not put a strain on the listener.
	Modified	**Natural speed of speech, with some pauses and reformulations that do not interrupt the flow.**
2 **B1.2**	**Current**	Noticeable pausing, false starts, reformulations and repetition.
	Modified	**Moderate speed of speech but interrupted by mid-clause pauses and reformulations.**
1 **B1.1**	**Current**	Noticeable pausing, false starts, reformulations and repetition.
	Modified	**Moderate speed of speech but interrupted by mid-clause pauses and reformulations.**
0 **A1/A2**	**Current**	Performance not sufficient for B1, or no meaningful language, or the responses are completely off-topic (memorised or guessing).

Table 8: Suggested fluency descriptors for Task 4

10.2 Recommendations for the Aptis Speaking training materials

The Aptis Speaking training materials were reviewed and evaluated in the light of the findings of this research. This section describes how the training materials have been analysed and offers our reflections, comments and suggestions for improvement.

10.2.1 Analysis of the Aptis Speaking training materials

Two of the authors independently took the Aptis online training for Aptis Speaking tests which was made available to them via the British Council. While the training materials for Speaking Tasks 1–4 were completed, we did not rate any of the available audio recordings from the candidates. Overall, we found the training materials very useful for a number of reasons. First, the Aptis Speaking test procedures were explained carefully and systematically. Second, there was a thorough discussion of the assessment process, the proficiency levels and the marking scheme each with useful examples. Finally, we found the training materials successful in providing the trainees with an overall understanding of the candidates' oral ability at different levels.

In our training sessions, we specifically focused on any notes about or discussions of fluency. The following are a list of the key observations we have made, and the emerging themes we have noticed in the materials. These are divided into three sections of *speed, breakdown,* and *repair* fluency.

Speed fluency

While a sense of speaking fast may be implicitly felt in the training materials, there were no explicit references to the speed with which candidates speak, e.g. speech rate or mean length of run between two obvious pauses. There are three references to rhythm and one to flow in the training materials. Rhythm is sometimes discussed in relation to pronunciation and sometimes to fluency. How rhythm is defined and measured is not discussed. Examples include:

- "Long pauses also add to a slow rhythm."
- "Pronunciation is clear with a natural-sounding rhythm."
- "Constant reformulation … and interrupts the flow."

Breakdown fluency

Silent pauses

Pausing is, in a large majority of examples in the training materials, the only reference made to indicate fluency. Given the important role of pausing as an indicator of one's fluency (as confirmed in this study), it is good that the training makes frequent references to this aspect of fluency. However, the materials do not discuss the quality of the pauses in terms of their length, frequency or location. Neither is pausing discussed in relation to different levels of proficiency or tasks. The most frequent references to pausing are:

- "There is frequent pausing."
- "There is some pausing but this does not affect fluency (or it does)."
- "There are some pauses in response 1 and 2."
- "There is some pausing but it is not overly noticeable."
- "There is some pausing at times but fluency is not affected. The candidate is able to keep going throughout the response."

Only in few places are there minor references to location or length of pauses. These are always abstract and indistinct, e.g. how long is long, or what is or is not noticeable.

- "There are frequent long pauses between words and phrases."
- "There is frequent pausing with long pausing at the beginning of each response."
- "There are longer pauses in response 3."

It was somehow alarming to see the instruction "*there is no pausing*", as trainees are left to make their own interpretation of this statement. Does this imply that "*no pausing*" is a characteristic of fluent speech? How would the trainees understand and interpret this?

It is good to see that occasionally the description of the pausing behaviour is slightly more detailed and the rationale is explained briefly.

- "There is some pausing throughout all 3 responses. Longer pauses are made in response 3 as the candidate struggles to search for vocabulary to express her ideas."

Filled pauses

There are no references to the use of non-lexically filled pauses in the training materials. As we listened to different test-takers' audio recordings, we observed that filled pauses were used more systematically by B and C level candidates (as compared to A2), and the use of filled pauses by C1 candidates seemed more effective in filling in mid-clause gaps in speech.

Repair fluency

Despite extensive references to repair features in all the three Aptis speaking scales (see Appendix 1), there are very few references to the repair aspects of fluency in the training materials. There is only one reference to hesitation (which might mean repairs as well as pauses), one to repetition and two to reformulation. These references do not clarify what these processes involve or how they affect performance.

- "There is some hesitation and pausing but fluency is generally fine."
- "There is some pausing and repetition of words and phrases."
- "Constant reformulation makes it very impeding and interrupts the flow."

Sometimes, repetition is discussed in relation to repeating the same concepts/vocabulary and therefore, it is not placed under fluency construct.

- "The candidate has sufficient basic vocabulary to respond to the questions, though lack of range results in responses being repetitive."

10.2.2 Recommendations

Speed measures are rather under-represented in the training materials. Given that speed fluency was found to be the most remarkable criterion feature to differentiate A2, B1 and B2 levels, it is highly recommended that the training materials should cover how raters can judge the speed aspect of candidates' speech.

Pausing seems to be at the heart of assessment of fluency. However, the training materials do not provide clear definitions for what pausing behaviour is deemed as satisfactory or effective at each level, which pausing behaviour is beneficial (filled versus silent pauses), and how pausing interacts with aspects of language production process. The final point requires further research before being fully incorporated in the training, but some interesting pausing behaviours were observed in the current research. They include pausing before reformulations, and before low-frequency, sophisticated lexical items.

Therefore, it is recommended that the training materials include information on:

- types of pause (silent vs filled pauses)
- which pausing behaviour is acceptable or interrupting understanding (mid-clause vs end-clause pauses)
- how pausing interacts with aspects of language production process (e.g. pausing before reformulations and sophisticated language).

Despite extensive references to repair features in the current rating scale, repair measures are rarely mentioned in the training materials. However, this makes sense in light of the results of the present study, which showed that repair measures are not straightforward to be used to differentiate levels. In addition, hesitations, replacements and repetitions do not always affect fluency in a negative way. Reformulation is a typical characteristic of speech production in both L1 and L2, and a moderate use of repair is not only natural but also a sign of aiming for more complex or accurate language. Different reformulation processes are at play in different levels of proficiency. Our initial observation suggests that quality of repair measures can indicate the level of proficiency, especially for C1 candidates. Although more research is clearly needed to examine this hypothesis more systematically, the effective use of reformulations at the C1 level is a possible aspect which the training materials could refer to.

11. CONCLUSIONS AND WAYS FORWARD

In order to contribute to enhancing the scoring validity of the Aptis Speaking test, this study has carried out a microanalysis of fluency features in candidates' output language at A2, B1, B2 and C1 levels. The analysis has identified criterial features in fluency at each level of proficiency, and it has also revealed the role of tasks in the assessment of fluency in the Aptis Speaking test. It is, therefore, believed that this research has offered a better understanding of the fluency construct measured by the Aptis Speaking test and provided fluency benchmarks at A2 to C1.

The empirical evidence offered in this study was then used to validate and/or to suggest recommendations to modify the Aptis Speaking test rating scales and rater training materials. From the outset of the study, the research team aimed at striking a balance between construct coverage and rater-usability (e.g. Taylor & Galaczi, 2011). That is, rating descriptors have to cover the measured construct as fully as possible, but they have to be short and succinct at the same time to be useful reference points to raters. Furthermore, to ensure the practical value of the research outcomes, a meeting was held with the British Council's Assessment Research Group (ARG) to discuss how best the outcomes of this research could inform possible revisions of the Aptis Speaking rating descriptors and rater training. It is hoped that these attempts were beneficial in bridging research in the field of Second Language Acquisition (SLA) and practice in language testing.

To extend the current study, four directions for future research can be suggested (some of which were discussed with the ARG team).

1) Microanalysis of other linguistic features elicited by the Aptis Speaking test

While this study focused solely on fluency, the Aptis Speaking holistic scale includes other assessment areas, such as topic relevance, grammatical range and accuracy, vocabulary range and accuracy, pronunciation, and cohesion. Some of these aspects have already been analysed closely. For example, Iwashita, May and Moore's (2017) recent mixed methods research provided valuable insights into differential performances of candidates elicited by different levels of proficiency in terms of vocabulary, coherence and cohesion. Their results have clear implications for those features of the Aptis rating scale. However, as far as the researchers are aware, not all aspects of spoken performance in Aptis have been scrutinised. Given the usefulness of the analysis for validating rating scales and understanding the construct measured (e.g. Brown, 2006a; Brown, Iwashita and McNamara, 2005; Nakatsuhara, 2014), it seems necessary to carry out detailed analyses of all linguistic and discoursal features that are designed to be measured in the Aptis Speaking test.

As the current study has already transcribed 128 task performances, some analyses of lexical and grammatical aspects can be relatively easily performed, using automated analysis tools such as TextInspector (http://www.textinspector.com/), Coh-Metrix (http://cohmetrix.com/) and Vocab Profile (http://www.lextutor.ca/vp/). Other analyses are likely to be more labour intensive (e.g. lexical and grammatical accuracy, pronunciation). The analysis of fluency in this study was indeed very labour-intensive and the sample size needed to be limited (i.e. 8 recordings x 4 levels x 4 tasks), which might have compromised the generalisability of the research outcomes. However, it is still believed that this study has offered useful indicators on how different fluency features are realised at different proficiency levels across different levels.

2) Rater perceptions of proficiency and the usefulness of the recommended rating descriptors

Based on various analyses of candidates' performances, this study has recommended some modifications to the fluency descriptors of the current Aptis Speaking scale. It is now highly important to explore the extent to which these empirically-informed fluency features are actually salient to Aptis raters. In other words, we need to confirm that trained raters can detect and use these features effectively in real time. Investigating raters' perceptions of proficiency when rating spoken performance has been demonstrated to be another useful method to develop and validate rating scales (e.g. Brown 2006b; Brown, Iwashita and McNamara, 2005; Ducasse & Brown 2009; May 2011; Orr 2002; Pollitt and Murray 1996). Indeed, Brown et al.'s (2005) TOEFL Speaking test study combined microanalyses of elicited linguistic features and a verbal report analysis of rater perceptions. Similarly, Brown's (2006a) investigation into candidates' linguistic features elicited in the IELTS Speaking test was complemented by another study that examined verbal reports produced by the IELTS examiners on the features they found salient while rating candidates' performances (Brown, 2006b).

3) Interactions between different linguistic/discoursal features and how a cluster of speech features is assessed

Another line of research is to examine how different features of speech interact with one other, and how a cluster of speaking features can be seen to distinguish candidates at different levels. As noted earlier, an interesting observation was made during the analysis of pauses in this study. Some pauses were located prior to reformulations, low-frequency lexical items, and sophisticated grammatical structures, indicating complex and variable interactions between different aspects of language. Such interactions have also been suggested by other researchers, such as Tonkyn (1999), Brown (2006a) and Seedhouse, Harris, Naeb and Üstünel (2014). For example, after a microanalysis of candidates' discourse in the IELTS Speaking test, Brown (2006a: 71) concluded that: "Overall, the findings indicate that while all the measures relating to one scale contribute in some way to the assessment on that scale, no one measure drives the rating; rather a range of performance features contribute to the overall impression of the candidate's proficiency". Although complex interactions of linguistic and discoursal features are not likely to be identified statistically in a meaningful way, a qualitative analysis of discourse as in Seedhouse et al. (2014) seems a promising method to uncover notable examples of interactions which can be shown in a rater training program. This appears to be particularly useful for a test that uses a holistic rating scale, like the Aptis Speaking test.

4) Comparing B2 and C1 candidates' performance in more demanding tasks

Finally, it is necessary to investigate whether and how fluency performances by B2 and C1 candidates can be differentiated. As noted earlier, a concern was raised in relation to the difficulty in differentiating B2 and C1 performances in terms of their fluency performance. While it could be due to a ceiling effect that comes into play at the B2 level for fluency aspects, it may suggest that the Aptis Speaking test, which lacks a C1 task, is not pushing B2 and C1 candidates to their linguistic limit for fluency. That is, the lack of a more demanding task at C1 might be preventing the test from eliciting differential performances from B2 and C1 candidates.

Therefore, it is recommended that future research should compare B2 and C1 candidates' performances using both B2 and C1 tasks. To this aim, it seems relevant to analyse how B2 and C1 candidates perform in the Aptis Advanced Speaking test which cover tasks targeting both B2 and C1 levels.

It is hoped that future research along these lines will further enhance the validity of the Aptis Speaking test. Lastly, while it is not often easy to integrate research findings to operational test designs, we believe our study has exemplified how ongoing validation studies could make tangible recommendations in a way that facilitates the test providers' modifications of operational tests.

REFERENCES

ACTFL (2014). American Council on the Teaching of Foreign Languages. Retrieved 1 November 2014 from: http://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

Ahmadian, M. (2012). The relationship between working memory capacity and L2 oral performance under task-based careful online planning condition. *TESOL Quarterly*, 46(1), 165–175.

Boersma, P. & Weenink, D. (2013). *Praat: doing phonetics by computer*. Version 5.3.51, retrieved 2 June 2013 from: http://www.praat.org/

Brown, A. (2006a). Candidate discourse in the revised IELTS Speaking Test*, IELTS Research Reports Vol 6*, 71–89. IELTS Australia, Canberra and British Council, London.

Brown, A. (2006b). An examination of the rating process in the revised IELTS Speaking Test, *IELTS Research Reports Vol 6*, 41–70. IELTS Australia, Canberra and British Council, London.

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 Series MS-29*, Educational Testing Service: Princeton, New Jersey.

Clark, H. & Fox Tree, J. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1), 73–111.

Council of Europe. (2014). *The Common European Framework*. Retrieved 10 February 2014 from: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

De Jong, N. H. & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In *The 6th Workshop on Disfluency in Spontaneous Speech (DiSS)* (pp. 17–20).

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

De Jong, N., Steinel, M., Florijn, A., Schoonen, R. & Hulstijn, J. (2012). Facets of Speaking Proficiency. *Studies in Second Language Acquisition*, 34(01), 5–34.

De Jong, N. & Perfetti, C. A. (2011). Fluency Training in the ESL Classroom: An experimental study of fluency development and proceduralization. *Language Learning*, 61(2), 533–568.

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. Approaches to second language acquisition. *Jyväskylä Cross-Language Studies*, 17, 119–133.

Ducasse, A. M. & Brown, A. (2011) The role of interactive communication in IELTS Speaking and its relationship to candidates' preparedness for study or training contexts, *IELTS Research Reports Vol 12*, 125–150. IDP: IELTS Australia, Canberra and British Council, London.

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. & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*(3), 299–323.

Foster, P. & P. Tavakoli (2009). Native speakers and task performance: Comparing effects on complexity, fluency and lexical diversity. *Language Learning***,** 59(4): 866–896.

Freed, B. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243–265). Ann Arbor: University of Michigan Press.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Longman.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Teting 13*(2), 208–238.

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 production. *IRAL*, 45, 215–240.

GCSE Modern foreign languages. (2014). New curriculum and subject content for the UK modern foreign languages. https://www.gov.uk/government/publications/gcse-modern-foreign-languages

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

Ishikawa, T. (2008). The effect of task demands of intentional reasoning on L2 speech performance. *The Journal of Asia TEFL, 5*(1), 29–63.

Iwashita, N., May, L. & Moore, P. (2017). Features of discourse and lexical richness at different performance levels in the APTIS speaking test, *ARAGs Research Reports Online*, AR-G/2017/2, 1–93.

Khang, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning.* 64(4): 809–854.

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

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

Levelt, W. J. M. (1989). *Speaking.* Cambridge, MA: Mit Press.

Mora, J. C. & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly*, 46(4), 610–641.

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* (pp. 141–74). Amsterdam: John Benjamins.

Nakatsuhara, F. (2012). The relationship between test-takers' listening proficiency and their performance on the IELTS Speaking Test. In L. Taylor & C. J. Weir (Eds.), IELTS Collected Papers 2: Research in reading and listening assessment. *Studies in Language Testing* vol. 34 (519–573). Cambridge: UCLES/CUP.

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

Orr, M. (2002). The FCE speaking test: using rater reports to help interpret test scores. *System* 30(2), 143–154.

O'Sullivan, B. & Dunlea, J. (2015*). Aptis Genearal Technical Manual Ver 1.0 TR/2015/005.* Available online at: www.britishcouncil.org/sites/default/files/aptis_general_technical_manual_v-1.0.pdf

Pollitt, A. & Murray, N. L. (1996) What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment. Selected papers from the 15th Language Testing Research Colloquium*, Cambridge and Arnhem. *Studies in Language Testing vol. 3* (74–91). Cambridge: UCLES/Cambridge University Press.

Prefontaine, Y. (2013) Perceptions of French fluency in second language speech production. *Canadian Modern Language Review* 69(3), 324–348.

Revesz, A., Ekiert, M. & Torgersen, E. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*. 1–22

Robinson, P. (2007). Task complexity. Theory of mind and intentional reasoning. *IRAL. 45*(2), 193–214.

Schmidt, M. S., and Fägersten, K. B. (2010). Disfluency markers in L1 attrition. *Language Learning, 60*(4), 753–791.

Seedhouse, P., Harris, A., Naeb, R. & Üstünel, E. (2014) The relationship between speaking features and band descriptors: A mixed methods study, *IELTS Research Reports Online Series* 2, 1–30. IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia.

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *IRAL* (Special Issue: New directions in L2 speech fluency). 54(2): 79–96.

Segalowitz, N. (2010). *The cognitive bases of second language fluency*. New York: Routledge.

Segalowitz, N. (2000). Automaticity and attentional skill in fluent performance. In H. Riggenbach (ed.), *Perspectives on Fluency*, 200–219. Ann Arbor: University of Michigan Press.

Skehan, P. (2014). The context for researching a processing perspective on task performance. In P. Skehan (Ed.), Processing Perspectives on Task Performance (Vol. 5, 1–26). Amsterdam: John Benjamins.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510–532.

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

Taylor, L. & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.) *Examining Speaking: Research and practice in assessing second language speaking* (pp.171–233). Cambridge: UCLES/CUP.

Tavakoli, P. (2016). Speech fluency in monologic and dialogic task performance. *IRAL (*Special Issue: New directions in L2 speech fluency). *54(2)*: 133–151

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers, *ELT Journal. 65*(1): 71–79.

Tavakoli, P. & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning. 58*(2): 439–473.

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

Tavakoli, P., Campbell, C. & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic Intervention. *TESOL Quarterly*. 50(2): 447–471.

Tonkyn, A. (1999). *Reading University/UCLES IELTS Rating Research Project – Interim Report.* Cambridge: internal UCLES Report.

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

APPENDIX 1:
Aptis Speaking rating scales

(Note: Fluency descriptors are highlighted in yellow.)

Speaking Task 1

Areas assessed: task fulfilment / topic relevance, grammatical range & accuracy, vocabulary range & accuracy, pronunciation, fluency.

5 **B1 (or above)**	Likely to be above A2 level.
4 **A2.2**	Responses to all **three** questions are on topic and show the following features • Some simple grammatical structures used correctly but basic mistakes systematically occur. • Vocabulary is sufficient to respond to the questions, although inappropriate lexical choices are noticeable. • Mispronunciations are noticeable and frequently place a strain on the listener. • Frequent pausing, false starts and reformulations but meaning is still clear.
3 **A2.1**	Responses to **two** questions are on topic and show the following features • Some simple grammatical structures used correctly but basic mistakes systematically occur. • Vocabulary is sufficient to respond to the questions, although inappropriate lexical choices are noticeable. • Mispronunciations are noticeable and frequently place a strain on the listener. • Frequent pausing, false starts and reformulations but meaning is still clear.
2 **A1.2**	Responses to at least **two** questions are on topic and show the following features • Grammatical structure is limited to words and phrases. Errors in basic patterns and simple grammar structures impede understanding. • Vocabulary is limited to very basic words related to personal information. • Pronunciation is mostly unintelligible except for isolated words. • Frequent pausing, false starts and reformulations impede understanding.
1 **A1.1**	Response to **one** question is on topic and shows the following features • Grammatical structure is limited to words and phrases. Errors in basic patterns and simple grammar structures impede understanding. • Vocabulary is limited to very basic words related to personal information. • Pronunciation is mostly unintelligible except for isolated words. • Frequent pausing, false starts and reformulations impede understanding.
0 **A0**	• No meaningful language or all responses are completely off-topic (e.g. memorised script, guessing).

Speaking Tasks 2 and 3

Areas assessed: task fulfilment / topic relevance, grammatical range & accuracy, vocabulary range & accuracy, pronunciation, fluency and cohesion.

5 **B2 (or above)**	Likely to be above B1 level.
4 **B1.2**	Responses to all **three** questions are on topic and show the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Sufficient range and control of vocabulary for the task. Errors occur when expressing complex thoughts. • Pronunciation is intelligible but inappropriate mispronunciations put an occasional strain on the listener. • Some pausing, false starts and reformulations. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
3 **B1.1**	Responses to **two** questions are on topic and show the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Sufficient range and control of vocabulary for the task. Errors occur when expressing complex thoughts. • Pronunciation is intelligible but inappropriate mispronunciations put an occasional strain on the listener. • Some pausing, false starts and reformulations. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
2 **A2.2**	Responses to at least **two** questions are on topic and show the following features • Uses some simple grammatical structures correctly but systematically makes basic mistakes. • Vocabulary will be limited to concrete topics and descriptions. Inappropriate lexical choices for the task are noticeable. • Mispronunciations are noticeable and put a strain on the listener. • Noticeable pausing, false starts and reformulations. • Cohesion between ideas is limited. Responses tend to be a list of points.
1 **A2.1**	Response to **one** question is on topic and shows the following features • Uses some simple grammatical structures correctly but systematically makes basic mistakes. • Vocabulary will be limited to concrete topics and descriptions. Inappropriate lexical choices for the task are noticeable. • Mispronunciations are noticeable and put a strain on the listener. • Noticeable pausing, false starts and reformulations. • Cohesion between ideas is limited. Responses tend to be a list of points.
0	• Performance below A2, or no meaningful language or the responses are completely off-topic (e.g. memorised script, guessing).

Speaking Task 4

Areas assessed: task fulfilment / topic relevance, grammatical range & accuracy, vocabulary range & accuracy, pronunciation, fluency and cohesion.

6 **C2**	Likely to be above C1 level.
5 **C1**	Response addresses all **three** questions and is well-structured. • Uses a range of complex grammar constructions accurately. Some minor errors occur but do not impede understanding. • Uses a range of vocabulary to discuss the topics required by the task. Some awkward usage or slightly inappropriate lexical choices. • Pronunciation is clearly intelligible. • Backtracking and reformulations do not fully interrupt the flow of speech. • A range of cohesive devices are used to clearly indicate the links between ideas.
4 **B2.2**	Responses to all **three** questions are on topic and show the following features • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. • Pronunciation is intelligible. Mispronunciations do not put a strain on the listener or lead to misunderstanding. • Some pausing while searching for vocabulary but this does not put a strain on the listener. • A limited number of cohesive devices are used to indicate the links between ideas.
3 **B2.1**	Responses to **two** questions are on topic and show the following features • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. • Pronunciation is intelligible. Mispronunciations do not put a strain on the listener or lead to misunderstanding. • Some pausing while searching for vocabulary but this does not put a strain on the listener. • A limited number of cohesive devices are used to indicate the links between ideas.
2 **B1.2**	Responses to at least **two** questions are on topic and show the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Limitations in vocabulary make it difficult to deal fully with the task. • Pronunciation is intelligible but occasional mispronunciations put an occasional strain on the listener. • Noticeable pausing, false starts, reformulations and repetition. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
1 **B1.1**	Response to **one** question is on topic and shows the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Limitations in vocabulary make it difficult to deal fully with the task. • Pronunciation is intelligible but occasional mispronunciations put an occasional strain on the listener. • Noticeable pausing, false starts, reformulations and repetition. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
0 **A1/A2**	Performance not sufficient for B1, or no meaningful language, or the responses are completely off-topic (memorised or guessing).

APPENDIX 2: Descriptive statistics

Table A2.1: Descriptive statistics for fluency measures across proficiency levels

		Mean	Std. Deviation	Minimum	Maximum
Speed Fluency					
Mean length of run pruned.1: Mean length of run pruned	**A2**	3.21	.73	2.28	4.82
	B1	5.84	1.88	3.55	10.59
	B2	8.54	1.91	4.09	11.83
	C1	7.75	1.80	4.20	12.18
	Total	6.54	2.58	2.28	12.18
Articulation rate pruned.1: Articulation rate pruned	**A2**	158.05	23.22	116.21	221.17
	B1	188.04	30.00	150.46	241.90
	B2	224.25	33.53	171.51	291.01
	C1	234.90	36.27	177.75	326.03
	Total	204.19	42.84	116.21	326.03
Speech rate pruned.1: Speech rate pruned	**A2**	73.24	18.35	37.64	103.52
	B1	135.21	31.57	97.67	195.90
	B2	172.06	25.81	127.29	214.76
	C1	172.18	35.51	110.85	245.23
	Total	142.50	47.75	37.64	245.23
Phonation time ratio.1: Phonation time ratio	**A2**	46.45	10.32	27.59	64.48
	B1	71.29	6.59	57.72	83.32
	B2	76.85	5.05	67.82	85.63
	C1	72.91	7.01	59.45	88.20
	Total	68.24	13.24	27.59	88.20
Total speaking time.1: Total speaking time	**A2**	35.78	14.96	13.62	67.89
	B1	77.08	15.46	43.00	104.12
	B2	86.09	15.84	54.49	108.07
	C1	82.20	15.58	49.58	108.23
	Total	72.59	24.22	13.62	108.23
Total sample time.1: Total sample time	**A2**	78.68	29.60	27.82	130.50
	B1	108.71	21.74	52.01	131.08
	B2	112.05	19.08	72.27	131.15
	C1	112.74	18.42	79.39	130.15
	Total	104.67	25.46	27.82	131.15
Repair Fluency					
Number of repetitions	**A2**	2.83	4.50	0	14
	B1	5.13	4.45	0	17
	B2	4.41	3.61	0	12
	C1	5.09	3.31	0	14
	Total	4.47	4.01	0	17

		Mean	Std.	Minimum	Maximum
Number of reformulations and false starts	**A2**	.83	1.37	0	5
	B1	2.78	2.17	0	9
	B2	2.06	1.79	0	7
	C1	1.75	1.39	0	6
	Total	1.93	1.84	0	9
Number of self-corrections	**A2**	.38	.65	0	2
	B1	1.34	1.21	0	4
	B2	1.16	1.72	0	8
	C1	1.22	1.58	0	7
	Total	1.07	1.42	0	8
Total repairfluency.1: Total repair fluency	**A2**	4.04	5.89	0	19
	B1	9.25	5.75	0	22
	B2	7.63	5.52	1	22
	C1	8.06	3.73	2	17
	Total	7.46	5.50	0	22
Breakdown Fluency					
Length of filled pauses	**A2**	.43	.33	.00	1.24
	B1	.53	.23	.00	1.08
	B2	.41	.26	.00	.84
	C1	.52	.15	.00	.78
	Total	.48	.25	.00	1.24
Length of silent pauses	**A2**	1.42	.67	.61	3.07
	B1	.63	.23	.34	1.43
	B2	.56	.13	.38	.94
	C1	.54	.12	.36	.94
	Total	.75	.47	.34	3.07
Length of total pauses mean	**A2**	1.61	.59	.93	2.95
	B1	.75	.17	.41	1.23
	B2	.71	.18	.46	1.11
	C1	.74	.16	.39	1.08
	Total	.91	.46	.39	2.95
Length of mid-clause pauses	**A2**	1.50	.67	.59	3.48
	B1	.69	.13	.41	.96
	B2	.62	.14	.39	1.01
	C1	.66	.15	.44	1.07
	Total	.82	.47	.39	3.48
Length of end-clause pauses	**A2**	1.70	.65	.97	3.37
	B1	.80	.25	.41	1.57
	B2	.75	.24	.46	1.30
	C1	.79	.19	.38	1.19
	Total	.97	.51	.38	3.37
Length of end silent pauses average	**A2**	1.55	.88	.48	3.70
	B1	.68	.30	.34	1.76
	B2	.60	.20	.39	1.32
	C1	.57	.15	.35	1.10
	Total	.80	.58	.34	3.70

		Mean	Std.	Minimum	Maximum
Length of end filled pauses average	**A2**	.36	.34	.00	1.24
	B1	.39	.31	.00	.86
	B2	.32	.29	.00	.84
	C1	.49	.26	.00	.89
	Total	.39	.30	.00	1.24
Length of mid silent pauses average	**A2**	1.25	.69	.00	2.82
	B1	.54	.18	.00	.86
	B2	.50	.12	.35	1.01
	C1	.45	.14	.00	.70
	Total	.65	.45	.00	2.82
Length of mid filled pauses average	**A2**	.30	.35	.00	.85
	B1	.47	.29	.00	1.08
	B2	.37	.27	.00	.83
	C1	.47	.23	.00	.92
	Total	.41	.29	.00	1.08
Number of total pauses pm	**A2**	21.40	5.22	12.70	31.42
	B1	23.16	3.73	13.93	30.48
	B2	19.99	3.68	13.80	30.41
	C1	22.00	2.91	17.22	27.65
	Total	21.66	4.00	12.70	31.42
Number of filled pauses pm.	**A2**	3.08	3.90	.00	17.34
	B1	7.83	7.83	.00	22.40
	B2	5.29	6.53	.00	28.29
	C1	8.61	5.60	.00	18.97
	Total	6.41	6.53	.00	28.29
Number of silent pauses pm.1	**A2**	29.11	9.87	8.75	44.73
	B1	28.63	13.05	1.39	48.72
	B2	24.70	10.35	6.44	44.21
	C1	23.65	10.58	5.78	50.44
	Total	26.35	11.22	1.39	50.44
Number of mid-clause pm.1	**A2**	21.47	8.44	7.80	38.12
	B1	21.79	4.80	10.15	29.26
	B2	15.03	6.07	4.04	29.70
	C1	18.72	6.58	10.07	31.21
	Total	19.10	6.94	4.04	38.12
Number of end-clause pm.	**A2**	21.33	7.19	11.81	40.69
	B1	24.47	6.26	13.84	41.33
	B2	24.90	6.28	13.81	35.32
	C1	25.26	4.16	16.05	30.42
	Total	24.17	6.09	11.81	41.33
Number of end silent pm.	**A2**	7.09	2.96	1.48	13.80
	B1	8.05	3.62	.70	15.48
	B2	8.37	3.96	2.30	15.59
	C1	7.49	3.41	1.93	14.07
	Total	7.79	3.53	.70	15.59

		Mean	Std.	Minimum	Maximum
Number of end filled pm.	**A2**	.87	.92	.00	3.25
	B1	1.55	1.96	.00	7.70
	B2	1.07	1.30	.00	3.77
	C1	1.59	1.18	.00	3.85
	Total	1.30	1.44	.00	7.70
Number of mid silent pm.	**A2**	7.47	3.77	.00	13.80
	B1	6.27	3.55	.00	12.40
	B2	3.98	2.39	.70	10.82
	C1	4.34	2.84	.00	11.16
	Total	5.38	3.40	.00	13.80
Number of mid filled pm.	**A2**	.67	1.18	.00	5.42
	B1	2.36	2.41	.00	7.59
	B2	1.57	2.16	.00	10.61
	C1	2.72	2.07	.00	7.30
	Total	1.91	2.16	.00	10.61

Table A2.2: Descriptive statistics for fluency measures across tasks

		Mean	Std. Deviation	Minimum	Maximum
Speed Fluency					
Mean length of run pruned.1: Mean length of run – pruned	**T1**	6.32	2.52	2.67	12.18
	T2	6.14	2.56	2.43	10.81
	T3	6.39	2.85	2.28	11.83
	T4	7.58	2.15	3.95	11.09
	Total	6.54	2.58	2.28	12.18
Articulation rate pruned.1: Articulation rate – pruned	**T1**	202.09	41.56	139.48	283.14
	T2	197.61	43.10	137.42	291.01
	T3	201.43	43.49	116.21	302.02
	T4	219.46	42.39	150.46	326.03
	Total	204.19	42.84	116.21	326.03
Speech rate pruned.1: Speech rate – pruned	**T1**	141.50	47.57	44.22	245.23
	T2	135.61	49.16	50.19	212.76
	T3	136.89	51.44	37.64	218.26
	T4	160.51	38.40	105.01	244.25
	Total	142.50	47.75	37.64	245.23
Phonation time ratio.1: Phonation time ratio	**T1**	68.50	13.41	29.62	88.20
	T2	66.89	14.30	33.02	83.32
	T3	65.97	15.30	27.59	83.16
	T4	72.70	6.58	57.72	85.63
	Total	68.24	13.24	27.59	88.20
Total speaking time.1: Total speaking time	**T1**	53.49	16.00	15.79	73.59
	T2	79.16	26.00	13.62	107.91
	T3	77.43	26.68	25.89	108.23
	T4	82.83	10.50	48.43	95.36
	Total	72.59	24.22	13.62	108.23

		Mean	Std. Deviation	Minimum	Maximum
Total sample time.1: Total sample time	**T1**	76.95	14.18	27.82	87.04
	T2	115.18	23.66	40.63	131.15
	T3	114.96	23.98	52.01	130.36
	T4	113.88	9.79	70.91	119.64
	Total	104.67	25.46	27.82	131.15
Repair Fluency					
Number of repetitions	**T1**	2.78	2.47	0	9
	T2	4.84	4.73	0	16
	T3	4.91	4.37	0	17
	T4	5.63	3.63	1	15
	Total	4.47	4.01	0	17
Number of reformulations and false starts	**T1**	1.44	1.34	0	5
	T2	1.97	1.80	0	6
	T3	2.41	2.12	0	9
	T4	1.88	1.98	0	7
	Total	1.93	1.84	0	9
Number of self-corrections	**T1**	.78	1.16	0	4
	T2	1.22	1.31	0	4
	T3	1.13	1.70	0	8
	T4	1.17	1.49	0	7
	Total	1.07	1.42	0	8
Total repairfluency.1: Total repair fluency	**T1**	5.00	3.72	0	15
	T2	8.03	6.22	0	22
	T3	8.44	5.85	0	22
	T4	8.67	5.26	1	20
	Total	7.46	5.50	0	22
Breakdown Fluency					
Length of filled pauses	**T1**	.45	.24	.00	.84
	T2	.46	.25	.00	.75
	T3	.53	.28	.00	1.24
	T4	.46	.23	.00	.76
	Total	.48	.25	.00	1.24
Length of silent pauses	**T1**	.73	.43	.34	1.95
	T2	.77	.44	.37	2.21
	T3	.86	.65	.40	3.07
	T4	.60	.20	.36	1.20
	Total	.75	.47	.34	3.07
Length of total pauses mean	**T1**	.88	.44	.39	2.32
	T2	.94	.46	.41	2.23
	T3	1.01	.60	.47	2.95
	T4	.77	.20	.45	1.23
	Total	.91	.46	.39	2.95
Length of mid-clause pauses	**T1**	.82	.43	.42	2.39
	T2	.86	.47	.40	2.38
	T3	.95	.61	.51	3.48
	T4	.62	.14	.39	.88
	Total	.82	.47	.39	3.48

		Mean	Std. Deviation	Minimum	Maximum
Length of end-clause pauses	**T1**	.95	.51	.38	2.75
	T2	.98	.52	.41	2.59
	T3	1.04	.64	.44	3.37
	T4	.88	.27	.47	1.57
	Total	.97	.51	.38	3.37
Length of end silent pauses average	**T1**	.82	.60	.34	2.67
	T2	.81	.53	.39	2.58
	T3	.89	.75	.39	3.70
	T4	.67	.27	.38	1.55
	Total	.80	.58	.34	3.70
Length of end filled pauses average	**T1**	.39	.29	.00	.89
	T2	.38	.28	.00	.76
	T3	.41	.35	.00	1.24
	T4	.37	.29	.00	.88
	Total	.39	.30	.00	1.24
Length of mid silent pause average	**T1**	.58	.43	.00	2.16
	T2	.69	.41	.34	2.01
	T3	.80	.59	.36	2.82
	T4	.48	.12	.35	.86
	Total	.65	.45	.00	2.82
Length of mid filled pause average	**T1**	.37	.29	.00	.84
	T2	.42	.30	.00	.92
	T3	.43	.30	.00	1.08
	T4	.40	.26	.00	.80
	Total	.41	.29	.00	1.08
Number of total pauses pm.	**T1**	22.07	4.45	12.70	30.48
	T2	21.80	3.70	13.80	30.25
	T3	21.24	4.49	13.81	31.42
	T4	21.46	3.18	16.72	28.93
	Total	21.66	4.00	12.70	31.42
Number of filled pauses pm.	**T1**	7.39	7.74	.00	28.29
	T2	5.65	5.64	.00	21.93
	T3	5.70	6.34	.00	21.81
	T4	7.08	6.29	.00	19.88
	Total	6.41	6.53	.00	28.29
Number of silent_pauses_pm.1	**T1**	26.28	11.35	1.39	43.62
	T2	27.94	11.21	7.63	44.21
	T3	25.49	11.75	3.79	50.44
	T4	25.47	10.82	6.28	48.72
	Total	26.35	11.22	1.39	50.44
Number of midclause_pm.1	**T1**	18.55	6.86	9.53	34.50
	T2	20.03	7.37	4.04	35.86
	T3	19.36	7.57	7.80	38.12
	T4	18.24	5.75	5.15	31.21
	Total	19.10	6.94	4.04	38.12
Number of end-clause pm.	**T1**	25.49	6.64	12.94	39.88
	T2	23.53	6.56	11.81	40.69
	T3	23.09	5.05	13.81	34.39

		Mean	Std. Deviation	Minimum	Maximum
	T4	24.68	5.93	14.40	41.33
	Total	24.17	6.09	11.81	41.33
Number of end silent pm.	**T1**	8.31	3.47	.70	15.48
	T2	7.79	3.47	1.48	15.59
	T3	7.30	3.78	1.42	14.07
	T4	7.78	3.46	1.05	12.73
	Total	7.79	3.53	.70	15.59
Number of end filled pm.	**T1**	1.51	1.77	.00	7.70
	T2	1.25	1.26	.00	4.29
	T3	1.17	1.30	.00	3.85
	T4	1.25	1.38	.00	5.23
	Total	1.30	1.44	.00	7.70
Number of mid silent pm.	**T1**	4.83	3.64	.00	12.94
	T2	6.19	3.69	.47	13.80
	T3	5.45	3.27	.47	12.50
	T4	4.96	2.76	1.04	11.67
	Total	5.38	3.40	.00	13.80
Number of mid filled pm.	**T1**	2.19	2.65	.00	10.61
	T2	1.57	1.82	.00	6.67
	T3	1.68	2.02	.00	7.59
	T4	2.29	2.07	.00	6.54
	Total	1.91	2.16	.00	10.61

APPENDIX 3: Glossary

This glossary provides a list of simple and operational descriptions for some of the technical terms used in the report. For ease of reference, it is organised alphabetically.

A – F

Breakdown fluency, i.e. the pauses and silences that break down the flow of speech

Composite pause: a pause that includes at least one silent and one filled pause

End-clause pause: a pause that occurs at the end of a clause

False start: abandoning of a word or linguistic unit that has just been uttered

Filled pause: a paused filled with non-lexical interjections such as hmm, uh, etc.

G – Q

Length of pause: how long an average pause is for each speaker in each task

Number of pauses: how frequently a speaker pauses during each task performance

Mid-clause pauses: a pause that occurs in the middle of a clause

Pause: a pause, whether filled or silent, which lasts at least 0.25 of a second (250 millisecond)

Pruning/pruned: pruning data involves excluding all the repetitions and hesitations

R – Z

Reformulation: modifying/reformulating a linguistic unit that has been uttered

Repair fluency, i.e. hesitations, repetitions and reformulations that are used to repair speech during the production process

Repetition: exact repetition of a word or phrase previously uttered

Speed fluency, i.e. speed with which speech is performed

Self-correction: the act of changing and reformulating a phrase or linguistic unit previously uttered for the purpose of correction

Silent pause: a pause where the pause is not interrupted by any sound

Total repair measures: the total number of the above measures, i.e. repetition, self-correction, false start and reformulation.

Utterance fluency refers to measurable aspects of fluency.

The list below shows how the different fluency measures were calculated.

Speed

- Total speaking time (time spent speaking with all pauses excluded)
- Total sample time (total amount of time to complete the task)
- Speech rate (pruned): total number of syllables divided by total time (including pauses) multiplied by 60
- Mean length of run (pruned): the mean number of syllables between two pauses
- Articulation rate (pruned): total number of syllables per minute divided by total amount of speaking time (excluding pauses) multiplied by 60

Breakdown

- Phonation time ratio: the ratio of total speaking time to total sample time (excluding pauses) given as a percentage
- Mean length of silent pauses per 60 seconds at mid-clause and end-clause positions
- Mean length of filled pauses per 60 seconds at mid-clause and end-clause positions
- Frequency of silent pauses per 60 seconds at mid-clause and end-clause positions
- Frequency of filled pauses per 60 seconds at mid-clause and end-clause positions
- Mean length of composite pauses per 60 seconds at mid-clause and end-clause positions
- Frequency of composite pauses per 60 seconds at mid-clause and end-clause positions

Repair measures

- Frequency of partial or complete repetitions (per 60 seconds)
- Frequency of self-corrections (per 60 seconds)
- Frequency of false starts and reformulations (per 60 seconds)
- Total number of repair measures used (per 60 seconds)

British Council Assessment Research Awards and Grants

If you're involved or work in research into assessment, then the British Council Assessment Research Awards and Grants might interest you.

These awards recognise achievement and innovation within the field of language assessment and form part of the British Council's extensive support of research activities across the world.

SCORING VALIDITY OF THE APTIS SPEAKING TEST: INVESTIGATING FLUENCY ACROSS TASKS AND LEVELS OF PROFICIENCY

AR-G/2017/7

Dr. Parvaneh Tavakoli
Dr. Fumiyo Nakatsuhara
Dr. Ann-Marie Hunter

ARAGs RESEARCH REPORTS ONLINE

ISSN 2057-5203