BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

# TOWARDS A MODEL OF MULTI-DIMENSIONAL PERFORMANCE OF C1 LEVEL SPEAKERS ASSESSED IN THE APTIS SPEAKING TEST

VS/2019/001

**Dr Fumiyo Nakatsuhara, CRELLA, University of Bedfordshire**
**Dr Parvaneh Tavakoli, University of Reading, UK**
**Dr Anas Awwad, Isra University, Jordan**

**BRITISH COUNCIL VALIDATION SERIES**
**SERIES EDITORS: RICHARD SPIBY AND CAROLYN WESTBROOK**
**ISSN 2398-7979**
**© BRITISH COUNCIL 2019**

# ABSTRACT

The current study draws on the findings of Tavakoli, Nakatsuhara and Hunter's (2017) quantitative study which failed to identify any statistically significant differences between various fluency features in speech produced by B2 and C1 level candidates in the Aptis Speaking test. This study set out to examine whether there were differences between other aspects of the speakers' performance at these two levels, in terms of lexical and syntactic complexity, accuracy and use of metadiscourse markers, that distinguish the two levels. In order to understand the relationship between fluency and these other aspects of performance, the study employed a mixed-methods approach to analysing the data. The quantitative analysis included descriptive statistics, *t*-tests and correlational analyses of the various linguistic measures. For the qualitative analysis, we used a discourse analysis approach to examining the pausing behaviour of the speakers in the context the pauses occurred in their speech.

The results indicated that the two proficiency levels were statistically different on measures of accuracy (weighted clause ratio) and lexical diversity (TTR and D), with the C1 level producing more accurate and lexically diverse output. The correlation analyses showed speed fluency was correlated positively with weighted clause ratio and negatively with length of clause. Speed fluency was also positively related to lexical diversity, but negatively linked with lexical errors. As for pauses, frequency of end-clause pauses was positively linked with length of AS-units. Mid-clause pauses also positively correlated with lexical diversity and use of discourse markers. Repair fluency correlated positively with length of clause, and negatively with weighted clause ratio. Repair measures were also negatively linked with number of errors per 100 words and metadiscourse marker type.

The qualitative analyses suggested that the pauses mainly occurred:

a) to facilitate access and retrieval of lexical and structural units
b) to reformulate units already produced
c) to improve communicative effectiveness.

A number of speech excerpts are presented to illustrate these examples.

It is hoped that the findings of this research offer a better understanding of the construct measured at B2 and C1 levels of the Aptis Speaking test, inform possible refinements of the Aptis Speaking rating scales, and enhance its rater training program for the two highest levels of the test.

# Authors


## Fumiyo Nakatsuhara

Fumiyo Nakatsuhara is a Reader in Language Assessment 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 speaking tests, task design and rating scale development. She has carried out a number of international testing projects, working with ministries, universities and examination boards. Fumiyo has recently published the book, *The Discourse of the IELTS Speaking Test* (with P. Seedhouse, 2018, CUP), and her work also appears in journals such as *Language Testing* (2011, 2014), *Language Assessment Quarterly* (2017) and *System* (2019).

## Parvaneh Tavakoli

Parvaneh Tavakoli is a Professor of Applied Linguistics at the University of Reading. Parvaneh's main research interest lies in the interface of second language acquisition, language teaching, and language testing. Parvaneh has led several international research projects investigating the effects of task and task design on performance, acquisition, assessment and policy in different contexts. Her research has been published in prestigious peer-reviewed journals and books.

## Anas Awwad

Anas Awwad is an Assistant Professor in Applied Linguistics at Isra University, Jordan. He was awarded his PhD in Applied Linguistics from University of Reading in 2017. His current research interests include second language acquisition, task-based language teaching, language testing and TESOL. His recent publications investigated the effects of task complexity on speech performance and the interaction between task complexity and learner individual differences.


# Acknowledgements

We would like to express our sincere gratitude to the British Council for funding and supporting this research. We are particularly grateful to Judith Fairbairn and Catherine Hughes for assisting us in obtaining Aptis recordings for Tavakoli, Nakatsuhara and Hunter (2017), which were re-analysed in this study. Our special thanks go to Dr Ann-Marie Hunter, who extracted relevant data from Tavakoli et al. (2017) and prepared them for the current research, as well as providing insightful comments at the analysis stage.

# CONTENTS

## APPENDICES



## LIST OF TABLES



## LIST OF FIGURES

# 1. INTRODUCTION

This research follows up a recently completed Aptis project titled 'Scoring validity of the Aptis Speaking Test: Investigating fluency across tasks and levels of proficiency' (Tavakoli, Nakatsuhara & Hunter, 2017). In order to contribute to enhancing the scoring validity of the Aptis Speaking Test, Tavakoli et al. (2017) 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. The research, therefore, 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 Tavakoli et al. was then used to validate and/or to suggest recommendations to modify the Aptis Speaking Test rating scales and rater training materials.

However, in Tavakoli et al. (2017), 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 distinguish A2 from B1, B1 from B2, the results failed to identify a useful measure to distinguish B2 and C1 performances. 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[1]. Indeed, Tavakoli et al.'s (2017) analysis made informal observations about interesting usage of pauses by high-level learners. C1 candidates seemed to pause before reformulations, low-frequency lexical items, and sophisticated grammatical structures, indicating complex and variable interactions between different aspects of language.

The current project is therefore aimed at helping develop a better understanding of fluency and its relationship to other aspects of performance at B2 and C1 level in the Aptis speaking test. The primary focus of the study is to explore what aspects of performance (i.e. lexical complexity, syntactic complexity, accuracy or discourse marker features) distinguish B2 and C1 levels of proficiency in the Aptis Speaking test, and in what ways a careful discourse-analysis approach to analysing speaking performance at higher proficiency levels can help develop an in-depth understanding of criterial features of performance. The findings of the study are expected to inform possible refinements of the Aptis Speaking Test scales and its rater training materials. In what follows, we will present an overview of the literature in this area before discussing the research design of the study and presenting the details of the analysis.

[1] 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. However, this is not within the scope of this project.

# 2. LITERATURE REVIEW

## 2.1 What is fluency?

Second language (L2) researchers commonly agree that although an important indicator of L2 proficiency and an effective means of examining L2 development, fluency is a complex and multifaceted construct which is often difficult to define and measure (Lennon, 1990; Segalowitz, 2010). Lennon (1990) argued that fluency is often used and understood in two different senses of 'broad' and 'narrow', where the former refers to 'mastery' of the language or general proficiency in it, the latter sense represented ease and automaticity of speech. Tavakoli and Hunter (2018) argued that fluency may represent four different but interrelated concepts that range from a broad sense of overall language proficiency to a very specific construct of temporal aspects of speech. They have also argued that these different definitions and representations may be useful for different purposes, with the very narrow definition, i.e. fluency in a microanalytic sense, being useful for professional purposes, such as language teaching and testing.

Fluency has been technically defined by a number of scholars. Koponen and Riggenbach (2000: 6) defined fluency as "flow, continuity, automaticity, or smoothness of speech", and Lennon (2000: 26) considered fluency as "rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention under the temporal constraints of on-line processing". As can be seen in these definitions, fluency is described with reference to its three key characteristics. First, fluency is defined through its observable features such as speed and pausing patterns. This aspect of fluency is perhaps the easiest to identify and observe since it focuses on concrete aspects of speech, such as length of pause and number of repetitions. The second characteristic which these definitions refer to is in relation to the impact of fluency on listeners, i.e. whether the listeners find it easy or difficult to understand and follow the speaker; for example, is speech disrupted too frequently or too slow to the extent that would put a strain on the listeners? The third characteristic used to explain fluency is the ease or complexity of the cognitive processes underlying the speech production process, i.e. how fluid and smooth the speech production process is.

An important point to note is that the three aspects of fluency are interrelated: when the processes involved in speech production occur smoothly and effortlessly, the output seems smoother and more uninterrupted which is, in turn, perceived as fluent by the listeners. These three aspects of fluency are termed as *utterance, perceived and cognitive* fluency in Segalowitz's (2010) triadic model.

As discussed earlier, given the observable nature of utterance fluency, this aspect of fluency has been the focus of many research projects, and as such our knowledge of utterance fluency has expanded over the past decades. Second language research (e.g. Skehan, 2003; Kormos, 2006; Kahng, 2014) has demonstrated that utterance fluency can best be measured in terms of speed (how fast someone speaks), breakdown (how much pausing is in her/his speech) and repair aspects (in what ways they repair their utterances for accuracy, appropriacy and better impact). It is also believed that each of the three aspects of fluency taps a different sub-construct of fluency (Tavakoli & Skehan, 2005), and may effectively contribute to a fuller understanding of how fluency is shaped. For a full review of measures of utterance fluency, see Tavakoli et al. (2017).

## 2.2 Fluency in language testing

While 'fluency' in speech can be conceptualised and defined in a number of ways in SLA research, fluency in language testing is usually operationalised in a tangible, readily measurable way, reflecting what Segalowitz (2010) calls *utterance fluency* and *perceived fluency*. The former concerns the quantifiable aspects of fluency such as speed, pauses and hesitation, and the latter relates to the impact on listeners and the inferences that listeners would make about a speaker's automaticity in speech production based on their perceptions of fluent speech. For example, in Hasselgreen's (2004) language testing study which looked at various aspects of spoken fluency, the operational definition of fluency informed by a comprehensive historical review of the construct of fluency was:

> "the ability to contribute to what a listener, proficient in the language, would normally perceive as coherent speech, which can be understood without undue strain, and is carried out at a comfortable pace, not being disjointed or disrupted by excessive hesitation" (Hasselgreen, 2004: 134).

In language assessment, fluency has consistently been one of the key constructs of L2 proficiency even in the earliest recorded tests of speaking, e.g. College Board's English Competence Examination (1930) (de Jong, 2018; Fulcher, 2003). Fluency is typically included in rating scales of most internationally recognised exams (e.g. Aptis, Cambridge General English exams, IELTS, Trinity ISE, TOEFL iBT, PTE Academic) and international language benchmarks for L2 communicative ability (e.g. ACTFL, CEFR). A careful examination of the role of fluency in such rating scales suggests that fluency is often combined with other aspects of performance, such as coherence in IELTS, and pronunciation and intelligibility in Trinity. Reviewing the construct of fluency across four international tests of speaking, de Jong (2018: 3) argued that:

> "Fluency is seen as a separate construct (PTEA); as a construct that goes hand in hand with pronunciation, on the one hand, and with complexity and accuracy, on the other (TOEFL iBT); as a construct that cannot be seen separately from coherence (IELTS); or as a construct that is part of the integral construct of language ability (ACTFL OPI)."

This brief analysis suggests that the rating scales in these different tests adopt a different interpretation of the construct of fluency, and consider a different set of interrelationships between fluency and other aspects of performance. This initial analysis, although rather surprising, can explain at least to some extent Brown's (2006) and Nakatsuhara's (2012) findings that suggested fluency as the most susceptible to task elicitation methods, and that raters tend to show the least confidence in evaluating it. Research in language assessment has provided ample evidence that fluency is directly linked with communicative adequacy (de Jong et al., 2015; Revesz et al., 2016), indicates and predicts proficiency (Iwashita et al., 2008; Revesz et al., 2016), and affects raters' perceptions of L2 ability (Prefontaine et al., 2016). Given the crucial role of fluency in language assessment, it is alarming that our knowledge of the construct of fluency seems rather limited.

While the importance of an evidence-based approach to developing and validating rating scales is frequently highlighted in the literature (e.g. Brown, 2006; Brown et al., 2005, Fulcher, 1996; Nakatsuhara, 2014), there are only a handful of studies focusing on fluency. Brown, et al. (2005), conducting a large-scale validation study on TOEFL, analysed 198 speech samples taken from test-takers of five proficiency levels. They analysed fluency of performances in terms of 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. Brown, et al. (20005) reported significant differences across different proficiency levels for speech rate, unfilled pauses, and total pause time, with medium or small effect sizes.

Nakatsuhara (2014) used a similar method to develop/validate new rating scales for the TEAP (Test of English for Academic Purposes) Speaking test. A small number of speech samples (N=23) across three proficiency levels were analysed for a range of different fluency measures including the number of unfilled pauses, total pause time, the ratio of repair, false starts and repetition, speech rate and articulation rate. Given the small sample size of the study, the use of inferential statistics was not possible. However, the results showed that the means of the three proficiency groups on all fluency measures varied in accordance with the rating scores, suggesting the linguistic analysis was a useful approach to validating the rating descriptors.

The final study we review here is Tavakoli et al. (2017) which the current project is aimed to extend. Working with 32 participants' performances across the four tasks of Aptis Speaking test, they analysed the data against a range of utterance fluency measures. Their analysis involved 19 measures in total (Tavakoli et al., 2017: 14) as noted below.

**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[2]

**Breakdown measures**

d) Phonation time ratio: percentage of performance time spent speaking
e) Mean length of all silent pauses
f) Mean length of silent pauses at mid-clause (f-1) and end-clause (f-2) positions, respectively
g) Mean length of filled pauses at mid-clause (g-1) and end-clause (g-2) positions, respectively
h) Frequency of all silent pauses
i) Frequency of silent pauses at mid-clause (j-1) and end-clause (j-2) positions, respectively
j) Frequency of filled pauses
k) Frequency of filled pauses at mid-clause (l-1) and end-clause (l-2) positions, respectively

**Repair measures**

l) Frequency of all repairs (per 60 seconds)
m) Frequency of false starts and reformulations (per 60 seconds)
n) Frequency of partial or complete repetitions (per 60 seconds)
o) Frequency of self-corrections (per 60 seconds)

As discussed earlier, the findings of Tavakoli et al. (2017) suggested that speed fluency distinguished A2, B1 and B2 levels, but B2 and C1 levels were not statistically different. Breakdown measures distinguished lower from higher levels, and repair measures did not systematically distinguish performance at different levels. Although it was not within the scope of the study, our informal observations indicated that the way in which pauses were used seemed to differ according to the level of proficiency. Tavakoli et al.'s (2017) study was particularly fruitful in suggesting possible modifications to the wording of the current Aptis fluency rating descriptors as presented in Appendix A.

[2] It should be noted that following de Jong et al. (2015), a pause is an unfilled silence of longer than 0.25 a second.

## 2.3 Multiple approaches to investigating fluency

The interest in researching fluency seems to be shared by a number of disciplines including second language acquisition and discourse analysis. Calling for more research on fluency in the context of language assessment, de Jong (2018) suggested that the four fields of applied linguistics, psycholinguistic, sociolinguistic and discourse analysis have important contributions to make to a full understanding of fluency. De Jong further argued that adopting a multi-disciplinary approach to analysing fluency, e.g. bringing a psycholinguistic and sociolinguistic perspective together in language assessment research projects "may help make oral language tests more valid, in particular if they lead to a more sophisticated manner of conceptualizing disfluencies" (de Jong, 2018: 14). Our current study is a response to this call. While the interest in researching fluency is rapidly growing among scholars from different disciplines including language assessment, there is only a limited number of studies that take a multidisciplinary approach to examining fluency. Tavakoli (2016) drew on the principles of discourse analysis to measure pauses in dialogic tasks. Working with L2 dialogic performances, Tavakoli (2016) argued that it is necessary to analyse interactive aspects of speech, e.g. interruptions, turn-taking and between-turn silence to develop a better understanding of fluency. This line of research coincides with McCarthy's (2010) notion of 'confluence', which indicates the co-construction of flow by more than one speaker in a conversation, and Nitta and Nakatsuhara's (2014) mixed-methods study with Conversational Analysis (CA) also demonstrated the difficulty in "determine[ing] the ownership of unfilled pause between turns in dialogues" (Nitta & Nakatsuhara, 2012: 155).

More recently, Seedhouse and Nakatsuhara's (2018) mixed-methods research explored how well the IELTS Speaking Test distinguish between Bands 5, 6, 7 and 8 in line with the IELTS band descriptors and which speaking features distinguish tests rated at these four levels. Using 60 transcripts of the IELTS Speaking Test, the former question was statistically investigated by quantifying selected features of constructs captured in the IELTS rating scales (e.g. fluency, grammatical complexity, range and accuracy), and the latter question was inductively explored from the spoken data, applying qualitative CA to transcripts of the speaking tests. Findings from their quantitative analysis echoed the findings of Brown's (2006: 71) earlier microanalysis of candidates' discourse in IELTS Speaking, which 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". Essentially, it seems that using such an atomic approach on its own to identifying which discrete individual components of a candidate's performance determine their score is not always informative. Seedhouse and Nakatsuhara's (2018) CA analysis demonstrated that statistical analysis on individual features should be complemented by a holistic perspective on test discourse, which can embrace complex interactions of various linguistic and discourse features. As such, a mixed-methods approach seems a promising method. A discourse analysis approach can not only help uncover notable examples of interactions to enable test designers and researchers to see a fuller picture of the relationship between candidates' output and scores, but also can offer useful examples to be shown in a rater training program to enhance raters' understanding of how different linguistic and discourse features could cluster in candidates' speech samples.

The research presented in the current report therefore pursued this line of research, by analysing how different features of speech interact with one another, and how a cluster of speaking features can be seen to distinguish candidates at different levels. This appears to be particularly useful for a test that uses a holistic rating scale, like the Aptis Speaking Test.

# 3. RESEARCH QUESTIONS

In order to better understand the criterial features that distinguish B2 and C1 performance elicited in the Aptis Speaking Test, this study's research questions were formulated as follows.

> **RQ1:** Are there any differences in how B2 and C1 candidates in the Aptis Speaking Test demonstrate their proficiency in terms of grammatical range and accuracy, vocabulary range and accuracy, and cohesion?
>
> **RQ2**: Do the above properties of B2 and C1 candidates interact with the fluency features identified in Tavakoli et al. (2017)?
>
> **RQ3:** Are there any differences between B2 and C1 candidates in the way pauses are used?

As detailed in Section 4.1 below, RQ1 and RQ2 were investigated through quantitative analyses on test-takers' output language, while RQ3 was explored through qualitative analyses.

# 4. METHODOLOGY

## 4.1. Mixed-methods approach

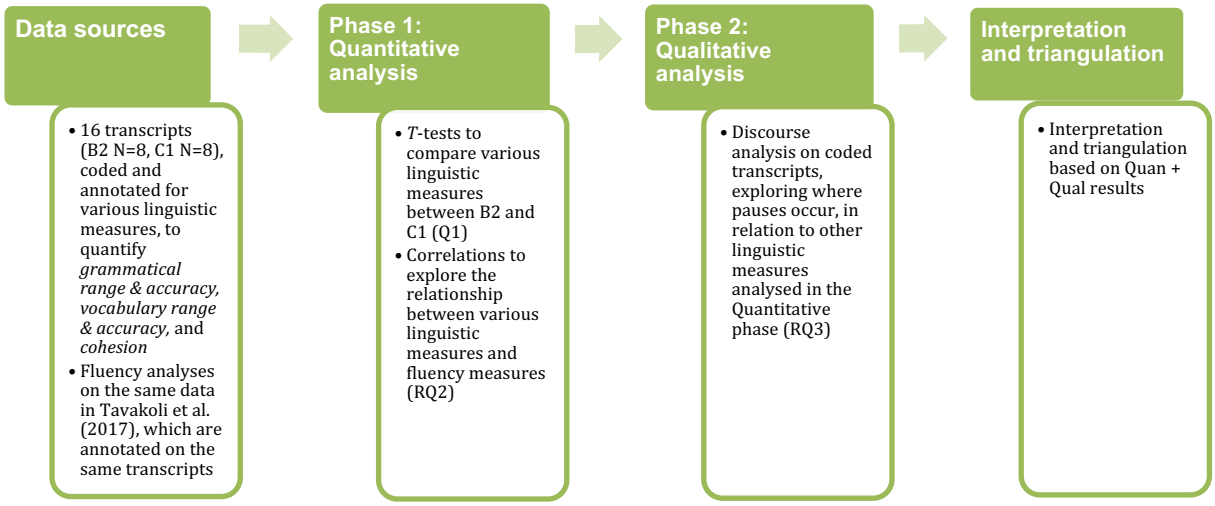
This study used a two-phased explanatory sequential mixed-methods design where qualitative data helps explain and build upon initial quantitative results (Creswell & Plano Clark, 2011). As further explained in Sections 4.2, 4.3 and 4.4 below, the quantitative and qualitative analyses shared the same data set, comprising 16 candidates' transcripts from Task 4 of the Aptis Speaking test. Quantitative analyses covered the first two research questions, investigating any differences between B2 and C1 candidates in terms of linguistic features measured for *grammatical range & accuracy, vocabulary range & accuracy,* and *cohesion* (RQ1), and examining the relationship between these features and the fluency features analysed on the same dataset in Tavakoli et al. (2017) (RQ2).

Following up on these results, the qualitative analysis in the second phase further scrutinised the coded transcripts around pauses (focusing particularly on long pauses), to explore the similarities and differences in the use of long pauses by B2 and C1 level candidates (RQ3).

Since the analysis for RQ3 does not touch upon all linguistic measures analysed in RQ1 and RQ2, the qualitative data are not fully applicable to interpret or triangulate quantitative results from the first research questions. However, wherever appropriate, RQ3 findings are used to interpret and triangulate the findings obtained for RQ1 and RQ2.

Figure 1 visually represents the mixed methods approach used in this study.

*Figure 1: Two-phased explanatory sequential mixed-method design in this study*




## 4.2. Materials

This study uses a subset of the data collected in Tavakoli et al. (2017). Of the 32 candidates' speech samples in their data set, 16 candidates' speech (eight B2 and eight C1 candidates) were selected for this study. Only Task 4 performances were examined in this study, since including other task performances would confound the findings of this study due to the target difficultly level of the other tasks. No new data collection was therefore undertaken.

The demographic information of the 16 candidates in this study, gathered through the operational Aptis test, is provided in Table 1. As shown in the second column, the IDs of B2 candidates start with B and those of C candidates commence with C. Candidates' L1s were only speculated based on the location of the test centre where they took the test, because the information is not gathered in the operational test.

Table 1 also shows the range of topics that the 16 candidates talked about. Since the audio recordings were taken from the operational Aptis test, it was unable for the researchers to obtain the actual topic prompts used in Task 4. Therefore, it should be noted that the topics listed in the table illustrate what the candidates talked about, which might be different from what the actual prompts required them to talk about. There were four topics that were observed more than once, and the remaining four topics were observed only once. Although we were unable to control topics for the choice of the 16 recordings, it seems that there was no noticeable bias between B2 and C1 recordings.

*Table 1: Demographic information of the 16 participants*

| Level | ID | Gender | Test centre location | Assumed L1 | Topic of Task 4 speech (from transcripts) |
|---|---|---|---|---|---|
| **B2** | B025 | M | Saudi Arabia | Arabic | Music produced by different cultures |
| **B2** | B028 | F | Kuwait | Arabic | My favourite piece of clothing |
| **B2** | B029 | F | Mexico | Spanish | The last time I visited an old building |
| **B2** | B030 | F | Colombia | Spanish | The last time I visited an old building |
| **B2** | B031 | M | Austria | German | The last time I watched a football match |
| **B2** | B032 | F | Uzbekistan | Uzbek | The last time I watched a football match |
| **B2** | B034 | M | Bosnia and Herzegovina | Bosnian | The last time I visited a tall building |
| **B2** | B035 | M | Georgia | Georgian | The last time I watched a football match |
| **C1** | C038 | F | India (Kerala) | Malayalam | The last time I helped someone |
| **C1** | C039 | F | Austria | German | My favourite piece of clothing |
| **C1** | C040 | F | Belgium | French | The last time I watched a football match |
| **C1** | C041 | F | Nigeria | English/local language | My best experience |
| **C1** | C044 | M | Georgia | Georgian | The last time I got lost |
| **C1** | C045 | M | Bangladesh | Bengali | My last long journey |
| **C1** | C047 | M | Colombia | Spanish | The last time I helped someone |
| **C1** | C048 | M | Ukraine | Ukrainian | The last time I watched a football match |

## 4.3. Quantitative analysis

### 4.3.1 Data coding

**The analytic measures**

The fundamental rationale for using analytic measures in this line of research is to quantify learners' L2 performance to enable researchers to "account for how and why language competencies develop for specific learners and target languages, in response to particular tasks, teaching, and other stimuli, and mapped against the details of developmental rate, route, and ultimate outcomes." (Norris & Ortega, 2009: 557). In this study, to carefully analyse how different features of each group's speech production interacted with one another, and how these features could discriminate candidates at different levels, a set of linguistic measures were used. Considering the Aptis scale descriptors and a line of research in SLA and Language Testing, several linguistic measures were selected to quantify the following features: *grammatical range and accuracy, vocabulary range and accuracy,* and *cohesion.* The previous studies that informed the selection of the measures include: Bax, Nakatsuhara and Wall (2019), Brown (2006), Brown, Iwashita and McNamara (2005), Foster and Wigglesworth (2016), Iwashita, Brown, McNamara and O'Hagan (2008), Iwashita, May and Moore (2017), Nakatsuhara (2014), Norris and Ortega (2009), Tavakoli and Foster (2008), and Tavakoli (2018).

The following analytic measures were adopted to capture each of the three linguistic features.

**1. Grammatical range and accuracy**

a) Number of verb elements per AS-unit
b) Syntactic complexity: mean length of AS-unit, mean length of clause, and ratio of subordination
c) Weighted clause ratio
d) Number of errors per 100 words

**2. Vocabulary range and accuracy**

a) Lexical sophistication
b) Lexical density
c) Lexical diversity
d) Number of lexical errors per 100 words

**3. Cohesion**

a) Types of metadiscourse markers
b) Tokens of metadiscourse markers

We did not examine pronunciation in this study since issues and challenges linked with pronunciation are more pertinent to the *articulation* stage of the language production process (Levelt, 1989) in which motor movements are executed to convert the linguistic plan into overt speech (Kormos, 2006). Research in this area suggests that some pronunciation problems might be linked to the process of encoding of pronunciation elements of the message as articulation (Skehan, 2014), and in the case of L2 speakers some might be linked to L1 transfer. Researchers have further argued that speech articulation is not believed "to be a major drain" to attention capacity and language processing demands (Bygate & Samuda, 2005). Given the limited scope of the current study, we were not able to conduct a careful cross-linguistic analysis of pronunciation challenges the test-takers from different L1 backgrounds may face, and pronunciation was therefore excluded from the proposed analysis.

The measures we chose to quantify *grammatical range and accuracy, vocabulary range and accuracy* and *cohesion* and the rationales of the selection are explained below.

#### 4.3.1.1 Measures of grammatical range and accuracy

To enable coding the data for the measures of grammatical range, the transcriptions were divided into syntactic units. The Analysis of Speech-unit (AS-unit), i.e. "a single speaker's utterance consisting of either independent clause or sub-clausal unit together with any subordinate clause(s) associated with either" (Foster, Tonkyn & Wigglesworth, 2000: 365) was adopted as the segmentation unit. The AS-unit was the favourite choice for its adequacy to deal with the messy nature of spoken data, unlike T-unit or C-unit (ibid), its popularity and reliability in SLA research (Norris & Ortega, 2009) and the straightforwardness of its codification (Ellis & Barkhuizen, 2005).

Grammatical range, known also as syntactic or structural complexity, is only one facet of the multi-componential construct of complexity. It can be defined as the quantity and quality of the distinct elements that a linguistic unit includes along with the relationship between these elements (Bulté & Housen, 2012). Producing a more complex and diverse grammatical range can also be deemed as a sign of language development and a trait of more proficient language users (Pallotti, 2009). Ellis and Barkhuizen (2005) define grammatical range as the extent to which a grammatical structure is varied and complex. Considering both the quality and quantity of linguistic structures as two facets of grammatical range, measuring syntactic complexity should therefore tap into the length, ratio, types and frequency of any clauses embedded in each linguistic unit which learners produce.

In this study, three measures were used to operationalise the multi-components of syntactic complexity, i.e. length, subclausal level and subordination (Norris & Ortega, 2009). The three measures were: a) mean length of AS-unit, i.e. the number of words divided by the number of AS-units; b) mean length of clause, i.e. the number of words divided by the number of clauses; and c) ratio of subordination, i.e. the number of clauses divided by the number of AS-units. Number of verb elements per AS-unit was another measure that was employed to tap into grammatical range. The choice of subordination-based, length-based and frequency-based measures of syntactic complexity corresponds to the notion that the three features are reliable in distinguishing speakers at higher levels of proficiency (Norris & Ortega, 2009). Furthermore, our choices respond to the calls that measuring syntactic complexity needs to take into account what the task requires in terms of the quality and quantity of subordinate clauses and grammatical structures (Inoue, 2016).

As for accuracy, it can be defined as the degree to which language performance is: i) error-free (Housen, Kuiken & Vedder, 2012); ii) attuned to a native language baseline (Yuan & Ellis, 2003); iii) deviant from a precise benchmark (Housen & Kuiken, 2009); iv) adequate and acceptable within a specific context (Pallotti, 2009); and v) affecting the flow of communication (Foster & Wigglesworth, 2016). Previous research on accuracy used: a) specific measures to quantify the use of definite linguistic elements (e.g. infinitives, past verbs, conjunctions ); b) global measures to monitor general accuracy (e.g. error-free clauses or units, number of errors); and c) error-gravity measures to count the effect of errors on communication or comprehensibility (Michel, 2017).

In this study, number of errors per 100 words (Mehnert, 1998) was employed as a global measure of accuracy since it is not affected by the coders' potential inconsistency in deciding clause and unit boundaries or by the spread of errors (Inoue, 2016), and for being more inclusive to all types of errors than other specific measures (Brown, Iwashita & McNamara, 2005). Weighted clause ratio which is an error-gravity measure (Foster & Wigglesworth, 2016) was also used as another measure of accuracy. Weighted clause ratio deals with errors based on their effect on communication and comprehensibility and hence sorts them according to their significance. This measure was considered in this study because it is fine-tuned to tackle the issues that traditional accuracy measures failed to satisfy, i.e. not crediting the correct language produced in erroneous clauses and not classifying errors based on their gravity. Foster and Wigglesworth suggest scoring 1.0 point for an accurate clause, .8 point for a clause that comprises any error that has no effect on the message, .5 point for any error that partially disturbs the message, and .1 point for any error that totally hampers the message. Weighted clause ratio is then calculated by dividing the total score by the total number of clauses in each sample.

#### 4.3.1.2 Measures of vocabulary range and accuracy

Several measures were employed to capture the aspects of lexical complexity and lexical accuracy. Lexical complexity, known also as vocabulary range, is that feature of language performance which indicates the speakers' tendency and ability to take the risk by using more complex and advanced lexis (Skehan, 2014). Lexical complexity in SLA is often defined in terms of lexical sophistication (e.g. considering frequency word lists), lexical density (e.g. the ratio of content and function words), and diversity of lexical items (e.g. type-token ratio).

In this study, lexical sophistication (Read, 2000) is calculated by the word frequency scores based on corpus-based frequency lists that cover BNC (K1, K2 and above, and Academic Word List) performed by the Vocabprofile (Cobb, 2013) function of Compleat Lexical Tutor (Cobb & Free, 2015). Lexical density is the ratio of content words (e.g., nouns, verbs, adjectives) to total number of words calculated by Vocabprofile (Cobb, 2013). To capture lexical diversity, two measures of type-token ratio (TTR) and D were adopted. TTR represents the range of different word types used in a text in relation to total number of word tokens (Richards, 1987). D is an advanced measure of TTR which corrects for the effect of any variation in text length on the results (Malvern & Richards, 2002). The use of D as an additional measure of TTR is justified by its appropriateness to gauge lexical diversity of tasks that include different contents or topics to elicit language performance (De Jong & Vercelloti, 2016). Text Inspector was used as an online software (Bax, 2012) for calculating the two measures of lexical diversity.

Number of lexical errors per 100 words was also included as a measure of lexical accuracy. A lexical error comprised any inaccurate word choice or nonnative-like selection of lexical chunks.

#### 4.3.1.3 Measures of cohesion

Cohesion refers to the presence of specific linguistic devices which can serve the purpose of linking a group of sentences to facilitate understanding (O'Reilly & McNamara, 2007). The use of appropriate metadiscourse markers is assumed to maintain cohesion in language performance since these linguistic devices signpost the various relationships (e.g., argumentation, conjunction, causal, addition) between the text or speech production elements (Schiffrin, Tannen & Hamilton, 2001). Whilst metadiscourse includes cohesion, it also extends beyond aspects of text/speech organisation, indicating an author's stance towards the content of the text/speech or towards the audience (Hyland, 2004: 109).

Recent research on metadiscourse markers in written texts suggest that while the overall use of metadiscourse markers increases as learners progress from novice to intermediate writers, the use of explicit markers decreases after reaching a certain level as they learn more sophisticated and subtle ways to express the organisation of a text without heavily depending on explicit markers (Bax et al., 2019). Indeed, Bax et al.'s (2019) study with B2-C2 learners showed that C2 learners used significantly fewer metadiscourse markers than C1 learners and that C1 learners used significantly fewer markers than B2 learners. However, when the variety of metadiscourse markers used by the three groups were examined, greater numbers of unique metadiscourse markers were employed as the level went up.

Therefore, focusing on the number and types of metadiscourse markers, this study examines whether or not the use of metadiscourse markers can be a distinguishing feature between B2 and C1 in the Aptis Speaking test. Text Inspector, a software tool for text analysis (Bax, 2012), was used to provide statistics on the range and tokens of the metadiscourse markers found in each transcript, i.e. *logical connectives, frame markers, attitude markers, relational markers, endophoric markers, sequence markers, personal markers, hedges, code glosses, emphatics and evidentials*.

Following coding the transcriptions and calculating all the measures by one researcher, all coded files were re-rated by another researcher to ensure the inter-rater reliability of the coding.

### 4.3.2 Statistical analyses

Descriptive statistics and independent samples *t*-tests were performed to examine whether there are any differences in the above features between B2 and C1 candidates (RQ1). Each comparison will also be visually represented. Pearson correlations were then used to explore the relationship between each of these linguistic features and the fluency features analysed in Tavakoli et al. (2017: RQ2). As noted in Section 2.2, the fluency measures analysed were the following:

**Speed measures**

1. Speech rate (pruned): total number of syllables divided by total performance time (including pauses) multiplied by 60.
2. Articulation rate: total number of syllables divided by total amount of phonation time (excluding pauses) multiplied by 60.
3. Mean length of run (pruned): the mean number of syllables between two pauses

**Breakdown measures**

4. Phonation time ratio: percentage of performance time spent speaking
5. Mean length of all silent pauses
6. Mean length of silent pauses at mid-clause (f-1) and end-clause (f-2) positions, respectively
7. Mean length of filled pauses at mid-clause (g-1) and end-clause (g-2) positions, respectively
8. Frequency of all silent pauses
9. Frequency of silent pauses at mid-clause (j-1) and end-clause (j-2) positions, respectively
10. Frequency of filled pauses
11. Frequency of filled pauses at mid-clause (l-1) and end-clause (l-2) positions, respectively

**Repair measures**

12. Frequency of all repairs (per 60 seconds)
13. Frequency of false starts and reformulations (per 60 seconds)
14. Frequency of partial or complete repetitions (per 60 seconds)
15. Frequency of self-corrections (per 60 seconds)

## 4.4 Qualitative analysis

The main aim of the qualitative analysis was to explore differences between B2 and C1 candidates in the use and placing of pause in their output language (RQ3).

While Seedhouse and Nakatsuhara (2018) employed Conversation Analysis (CA) to identify how a cluster of speech is constructing high-level performance (see Section 2.3), a limitation of CA is its subjective interpretation process. Combining the information obtained from linguistic analyses, the qualitative analysis of this study aimed to be as systematic and transparent as possible.

In the description of our qualitative analysis below, we consider our analysis approach as CA-informed discourse analysis[3]. CA is a useful paradigm of research to investigate and describe how talk-in-interaction is organised in ordinary conversations (e.g. Sacks, Schegloff & Jefferson, 1974) or in institutional talk (e.g. Heritage, 1997). In CA, it is of paramount importance to follow the principle of *unmotivated looking*, in other words, being open to discovering patterns or phenomena observed in the conversation *per se* without considering contextual features or speakers' identity. For most CA researchers, quantification of elements in talk is considered inappropriate unless these features are very well-defined and used to achieve a relatively limited range of communication goals (see Heritage, 1995; Schegloff, 1993). Over the years, the CA methodology has been applied to various types of discourse including L2 learners' speech in speaking assessments (e.g. Lazaraton, 2002; Seedhouse & Nakatsuhara, 2018) and, as such, variations of CA have derived, such as *Applied CA* (Richards & Seedhouse, 2005). However, it is still quite rare that CA is used with other types of linguistic or discourse analytic technique. As an innovative attempt to integrate and triangulate the results from our quantitative and qualitative analysis approaches in this mixed-methods study, we carried out CA on the transcripts that also denote various linguistic features analysed in the quantitative part of the research. This way, we hope that a more systematic and transparent description of learner language could be achieved, and that potential subjectivity involved in our CA interpretations could be minimised.

[3] Here, the term 'discourse analysis' serves as "an overall blanket term for any and all efforts to analyse "discourse"" (Have, 2006: 2), although within the UK academic context, it is often used in a rather specific way to indicate a particular research tradition in social psychology (for more discussions, see Have (2006) and Wooffitt (2005)).

The qualitative analysis proceeded in two phases.

***Phase 1: Preparing transcripts***

First, a number of measures identified in the quantitative analysis were mapped on a single transcript per candidate (see Figure 2). The annotations include: boundaries of AS units and clauses, verb elements, global errors, lexical errors, error free clauses, levels of errors (1-3), pauses, pause locations (mid clause or end of clause) (see Appendix B for coding symbols). This transcript was followed by vocabulary frequency lists according to the British National Corpus (BNC) and a list of academic words (Figure 3). Then, a list of the metadiscourse markers employed in the speech as well as another transcript that specifies the location of each metadiscourse marker was produced (Figure 4). This was carefully carried out by one of the researchers, and another researcher went through all transcripts to confirm the accuracy of the annotations.

*Figure 2: Transcript with annotations*

| Participant | Level | Task | No. of words |
|---|---|---|---|
| **25** | **B2** | **4** | **195** |

| | | | | | |
|---|---|---|---|---|---|
| No of AS-units \| | **16** | No of Clauses :: | **32** | No of verb elements | **32** |
| No of errors ≠ | **16** | No of lexical errors ≠~ | **9** | Error-free clauses errfr | **16** |
| Level **1** error (.8) | **8** | Level **2** error (.5) | **8** | Level **3** error (.1) | **0** |

our country (0.35mv) produces a wide variety of music (0.82ec) for from classical to pop to rock errfr (0.58eu) | and music is very & very (2.76mc ) very much in & in the people's lives ≠1 (0.82mv) | and the ≠1 most of our movies coming from our country (0.6eu) :: has ≠1 songs and dances :: featured in them errfr (0.57eu) | and these are very popular among (0.62mc) among & people errfr (2.08ec) | we do have an ≠1 classical music | we & we do ≠~2 a wide variety of (0.94mv) music and dances | also (1.63ec) i feel I (0.35mu) feel & very relaxed errfr :: when i (0.63mv) hear music ≠~2 | whenever i want errfr :: to concentrate on some work errfr :: i & i (0.33mu) listen to music errfr :: so i can concentrate on it better errfr (1.36eu) | and when i go on long drives errfr :: also i like ≠1 :: to hear ≠~2 (0.33mu) soothing music | so it keeps me errfr :: concentrated on & on the drive errfr (3.19ec) | it it & helps me errfr :: especially when i & when ≠~2 i'm working (0.68mv) on something very important (0.83eu) | to keep every ≠1 other noises and all this :: i put on music ≠1 :: and ≠~2 can (0.46mu) concentrate on my work (1.16eu) <clears throat> (0.73eu) | i think some cultures errfr :: produce so much music errfr :: because it is in their genes errfr | and they & they allow (1.43mc) music so much ≠~2 (2.04ec) | and (0.46ev) then they ≠~2 the freedom to produce these musics | the (1.77mc) wide variety of instruments which they have errfr (0.89eu) :: who give them & (4.10eu) who ≠~2 give them (1.26mc) this (0.89eu) :: to produce this ≠~1 so much music |

*Figure 3: Frequency lists and AWL*

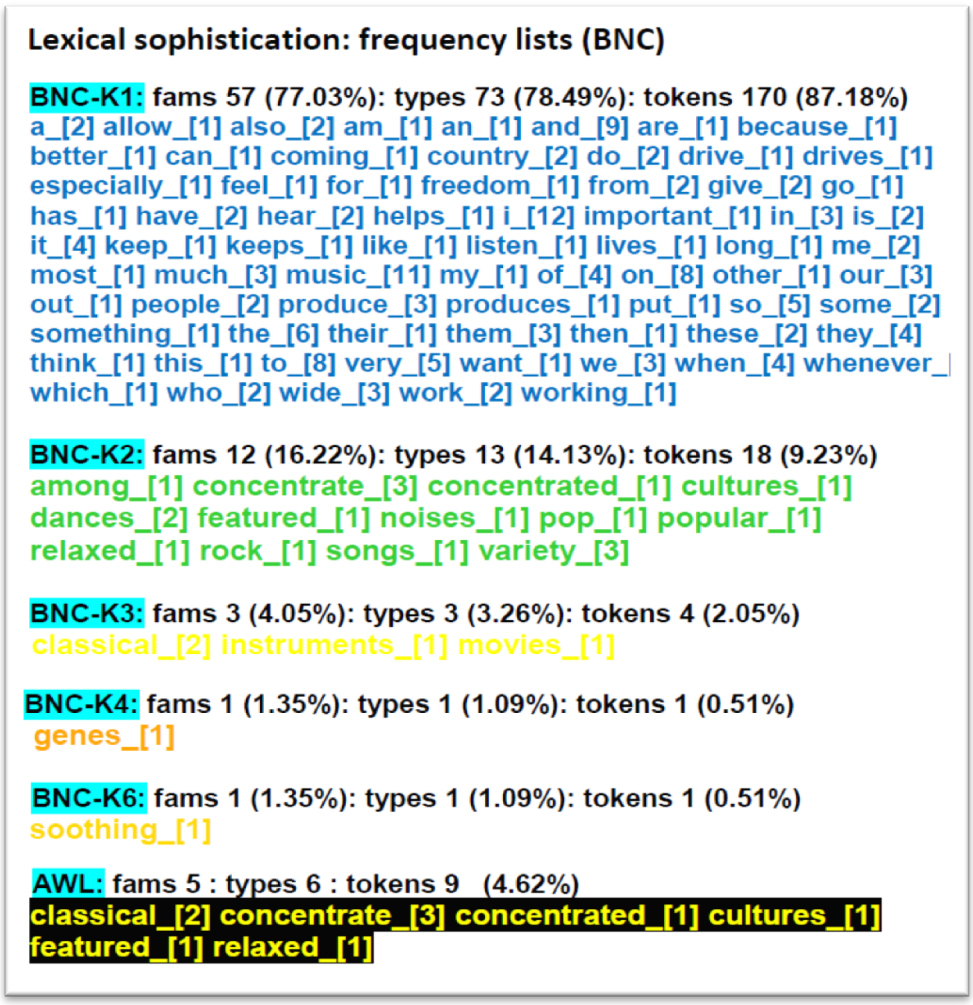


*Figure 4: List and location of metadiscourse markers*

**Metadiscourse Markers**

| | |
|---|---|
| Attitude marker | important (1) |
| Logical connective | also (2), and (9), because (1), so (2) |
| Person marker | I (12), me (2), my (1), our (3), we (3) |

our Person marker country produces a wide variety of music for from classical to pop to rock
and Logical connective music is very very in the people's lives and Logical connective the most of
our Person marker movies coming from our Person marker country has songs and Logical connective dances featured
in them and Logical connective these are very popular among people we Person marker do have an classical
music we Person marker we Person marker do a wide variety of music and Logical connective dances
also Logical connective i Person marker feel very relaxed when i Person marker hear music whenever i Person marker
want to concentrate on some work i Person marker i Person marker listen to music so Logical connective
i Person marker can concentrate on it better and Logical connective when i Person marker go on long drives
also Logical connective i Person marker like to hear soothing music so Logical connective it keeps me Person marker
concentrated on on the drive it helps me Person marker especially when i Person marker when i Person marker
am working on something very important Attitude marker to keep out the other noises i Person marker put
on music and Logical connective concentrate on my Person marker work i Person marker think some cultures
produce so much music because Logical connective it is in their genes and Logical connective they they allow
music so much and Logical connective then they the freedom to produce these musics the wide variety
of instruments which they have who give them who give them this to produce so much music

***Phase 2: Identifying emerged features***

In Phase 2, three researchers went through each transcript independently while taking notes, to investigate whether the linguistic features analysed in this study appear to interact with the fluency features identified in Tavakoli et al. (2017). To ensure to identify salient instances, it was decided to start the analysis with pauses longer than 1.0 second and then to move on to pauses shorter than 1.0 second. In addition, as we analysed, it became clearer that some mid-clause and end-clause pauses have differing functions. This is in line with the findings of Tavakoli et al.'s (2017) as well as other previous research (e.g. Tavakoli, 2011) that pausing at a mid-clause location is the sort of pause noticed by listeners and is one of the characteristics that could differentiate higher and lower-level L2 learners. In Tavakoli et al. (2017), the frequency of mid-clause silent pauses was significantly different between A2–B1 learners and B2–C1 learners. Therefore, for clarity all pauses longer than 1.0 second have been highlighted and colour coded in transcripts as in Figure 2 (mid-clause pauses in yellow and end-clause pauses in green).

Transcripts have then been examined by three researchers to identify salient types of utterance environment where long pauses occurred. As the three researchers examined the transcripts, they took notes of emerged themes and collected example excerpts. Given the small size of the dataset (N=16), the use of qualitative analysis software was not necessary. Instead, once all transcripts had been independently analysed, the three researchers met and compared emerged themes. One of the three researchers took the lead in the discussion, and all main and sub themes emerged, as well as each utterance example classified under different themes, were carefully examined altogether to reach a full consensus (see a summary table in Table 3).

During the identification of emerged themes and consensus building among the three researchers, the research team had the quantitative findings of RQ2 in mind, so that whenever possible, the qualitative data could help the interpretation and elaboration of the quantitative results. Therefore, the quantitative findings informed to some extent an initial list of themes and sub-themes, in line with Yin's (2011) suggestions on forming provisional categories. However, the qualitative analysis was also carried inductively, developing main and sub-themes in response to emergent aspects in the discourse data.

# 5. QUANTITATIVE RESULTS

The results section is structured around the study's research questions.

## 5.1 Research question 1

**RQ1. Are there any differences in how B2 and C1 candidates in the Aptis Speaking Test demonstrate their proficiency in terms of grammatical range and accuracy, vocabulary range and accuracy, and cohesion?**

Descriptive statistics and independent samples *t*-tests were performed to examine whether there were any statistically significant differences in the measures of complexity, accuracy and lexical variety between B2 and C1 level candidates. The means and standard variations, as well as the *t*-test results, are presented in Table 2 below. Where a significant result is obtained, Cohen's *d* is presented to show its effect size (highlighted in red type). Given the small sample size of the study, the *t*-test results should be interpreted with caution. However, it is believed that the labour-intensive close analyses we are conducting on various features will offer a comprehensible picture of overall performance characteristics by B2 and C1 candidates.

*Table 2: Descriptive and t-test results for B2-C1 levels*

| Measures | Level | Mean | SD | *t* | Sig. (2-tailed) | *d* |
|---|---|---|---|---|---|---|
| **Verb/AS-U** | B2 | 1.45 | .33 | -.131 | .894 | - |
| | C1 | 1.47 | .22 | | | |
| **Length/AS-U** | B2 | 11.94 | 1.96 | .013 | .989 | - |
| | C1 | 11.95 | 1.22 | | | |
| **Length/clause** | B2 | 6.85 | 1.43 | .472 | .666 | - |
| | C1 | 6.57 | .79 | | | |
| **Ratio of subordination** | B2 | 1.77 | .28 | -.441 | .654 | - |
| | C1 | 1.83 | .27 | | | |
| **Weighted Clause Ratio (WCR)** | B2 | .86 | .064 | -2.62 | .04 | 1.26 |
| | C1 | .92 | .031 | | | |
| **Error/100w** | B2 | 5.64 | 2.17 | 1.45 | .169 | - |
| | C1 | 4.35 | 1.32 | | | |
| **Lexical-error/100w** | B2 | 3.07 | 1.10 | 3.260 | .006 | 1.67 |
| | C1 | 1.52 | .714 | | | |
| **TTR** | B2 | .42 | .035 | -2.796 | .01 | 1.69 |
| | C1 | .48 | .047 | | | |
| **D** | B2 | 56.65 | 6.89 | -3.567 | .003 | 1.84 |
| | C1 | 73.23 | 10.72 | | | |
| **Lexical density** | B2 | .43 | .026 | -.083 | .994 | - |
| | C1 | .43 | .031 | | | |
| **K1 list** | B2 | 90.50 | 4.19 | .524 | .608 | - |
| | C1 | 89.28 | 5.11 | | | |
| **K2 list** | B2 | 5.45 | 3.50 | .108 | .916 | - |
| | C1 | 5.29 | 2.71 | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **K3 list** | B2 | 1.45 | 1.25 | .978 | .345 | - |
| | C1 | 2.20 | 1.79 | | | |
| **Measures** | **Level** | **Mean** | **SD** | ***t*** | **Sig. (2-tailed)** | ***d*** |
| **K4 and above list** | B2 | 1.66 | .80 | .819 | .427 | - |
| | C1 | 2.46 | 2.64 | | | |
| **AWL** | B2 | 1.73 | 1.61 | .010 | .992 | - |
| | C1 | 1.73 | .81 | | | |
| **Metadiscourse-marker type** | B2 | 16.38 | 3.77 | -1.485 | .160 | - |
| | C1 | 19.50 | 4.38 | | | |
| **Metadiscourse-marker token** | B2 | 55.75 | 19.69 | .216 | .832 | - |
| | C1 | 53.88 | 15.09 | | | |

As indicated in Table 2, the descriptive analysis suggests that the measures of lexical and syntactic complexity and accuracy are either higher for all measures in the C1 group or very similar between the two proficiency groups. However, the differences reach a significant level only for four measures, i.e., two measures of accuracy and two measures of lexical diversity. It is worth noting that the standard deviations are mostly small, suggesting that the participants in each group had similar linguistic abilities.

**Syntactic complexity**
With regard to syntactic complexity measures, there were no statistical differences between the two groups with regard to the number of verbs per AS-units, ratio of subordination, length of AS-units or length of clause. This suggests that at higher proficiency levels, there are not key differences between the syntactic structures in terms of range and variety the two proficiency levels produce.

**Accuracy**
Some statistically significant differences were observed when comparing the accuracy of performances of the two groups, highlighting the potential differences between the two proficiency levels. For the global measure of accuracy, i.e. weighted clause ratio (WRC), the results showed that C1 level produced more accurate clauses ($t$= 2.62, $p$ = .04, $d$ = 1.26). A statistically significant difference was also observed between the two groups for the number of lexical errors per 100 words ($t$= 3.26, $p$ = .006, $d$ = 1.67), but number of errors per 100 words did not reveal any significant differences.

**Lexical complexity**
The analysis of lexical complexity, measured in terms of lexical diversity, lexical density and lexical sophistication, demonstrated that the C1 group was statistically different from B2 for the two measures of lexical diversity, i.e., TTR ($t$= 2.79, $p$ = .01, $d$ = 1.69), and D ($t$= 3.56, $p$ = .003, $d$ = 1.84). However, there were no statistically meaningful differences between the two levels in terms of lexical density or lexical sophistication (frequency lists).

**Metadiscourse markers**
There were no statistically meaningful differences between the two groups in terms of their use of metadiscourse markers' type or token.

In sum, the results of the *t*-tests suggest that performance in the two levels of proficiency was different in terms of accuracy and lexical diversity. It is worth mentioning that, despite the small sample size of the study, the effect sizes obtained for the significant results, i.e., ranging between 1.26 and 1.84, are relatively large based on Polanski and Oswald's (2014) interpretations.

## 5.2 Research question 2

**RQ2. Do the above properties of B2 and C1 candidates interact with the fluency features identified in Tavakoli et al. (2017)?**

In order to investigate whether lexical and syntactic complexity and accuracy measures interacted with fluency measures studied in Tavakoli et al. (2017), a number of bivariate correlations were run. The results indicated several correlations between different aspects of fluency, i.e. speed, breakdown and repair measures. However, these are not reported here as they are not relevant to the focus of the current study (please see a detailed discussion of these relationships in Tavakoli et al., 2017). The correlations discussed here examine the relationship between measures of fluency on the one hand ,and measures of syntactic complexity, accuracy, lexical complexity and metadiscourse type and token on the other.

**Syntactic complexity**

The only measure of syntactic complexity that correlated with several measures of fluency was length of clause. It negatively correlated with mean length of run ($r$ = -.523, $p$ = .04), articulation rate ($r$ = -.651, $p$ < .006), and speech rate ($r$ = -.524, $p$ = .04), suggesting that longer clauses were produced more slowly. Length of clause positively correlated with total repair ($r$ = .610, $p$ = .01). interestingly, no correlations were observed between length of clause and breakdown measures of fluency. No significant correlations were found between measures of ratio of subordination or number of verbs per AS-unit. Only one significantly positive correlation was observed between length of AS-unit and frequency of end-clause pauses ($r$ = .527, $p$ = .04), implying that longer AS-units were associated with more end-clause pauses. Measures of syntactic complexity also correlated with one another and with measures of accuracy, but given the focus of the study, these are not discussed in this report.

**Accuracy**

As for accuracy measures, weighted clause ratio positively correlated with articulation rate ($r$ = .615, $p$ = .01), and negatively with total repair ($r$ = -.646, $p$ = .007), meaning more accurate performances were produced faster and had fewer repairs. This measure of accuracy also negatively correlated with length of clause ($r$ = -.657, $p$ = .006) and number of lexical errors per 100 words ($r$ = -.883, $p$ < .001). These two sets of correlations combined suggest that speakers with higher speed tend to produce shorter but more accurate clauses. Although lexical errors negatively correlated with WCR and D, there were no significant correlations between lexical errors and fluency measures. Number of errors per 100 words correlated positively with total repair ($r$ = -.614, $p$ = .01), and negatively with articulation rate ($r$ = -.605, $p$ < .01). Lexical errors per 100 words also negatively correlated with articulation rate ($r$ = -.531, $p$ = .03). The other correlations between accuracy and fluency measures did not reach a significant level.

**Lexical complexity**

The correlation analysis for measures of lexical complexity showed interesting results for the three types of complexity, i.e., diversity, density and sophistication. As for measures of lexical diversity, the results showed that D positively correlated with articulation rate ($r$ = .565, $p$ = .02), speech rate ($r$ = .503, $p$ = .05), and negatively correlated with lexical errors ($r$ = -.684, $p$ = .003). TTR positively correlated with frequency of mid-clause pauses ($r$ = .540, $p$ = .03), and length of mid-clause pauses ($r$ = .528, $p$ < .04). TTR also positively correlated with frequency of mid-clause filled pauses ($r$ = .676, $p$ = .004), suggesting that making mid-clause filled pauses was linked with levels of lexical diversity. As for measures of lexical sophistication, one significant positive correlation was obtained between K3 and length of mid-clause pauses ($r$ = .557, $p$ = .025), implying performances containing words from this were associated with more mid-clause pauses. There were no correlations between lexical density and fluency measures.

**Metadiscourse markers**
The most interesting correlational patterns emerged for the relationship between metadiscourse-marker measures and measures of fluency. Metadiscourse-marker type positively correlated with articulation rate ($r = .839$, $p < .001$), speech rate ($r = .798$, $p < .001$), and frequency of end-clause silent pauses ($r = .700$, $p = .003$). This measure negatively correlated with number of repetitions ($r = -.602$, $p = .01$), suggesting those who make more repetitions tend to produce fewer discourse-marker types.

The metadiscourse marker token measure positively correlated with mean length of run ($r = .618$, $p = .01$), articulation rate ($r = .681$, $p = .004$), speech rate ($r = .634$, $p < .008$), and frequency of end-clause pauses ($r = .697$, $p = .003$). This measure also negatively correlated with length of mid-clause pauses ($r = .582$, $p = .02$), and frequency of mid-clauses pauses ($r = .698$, $p = .003$). Some of the correlations reported for metadiscourse markers are the strongest relationships observed between the complexity, accuracy and lexis measures and fluency indices in the current study, suggesting the use of metadiscourse-markers may be closely linked with different aspects of fluency.

## 5.3 Summary of the correlations

**Syntactic complexity:** Length of clause negatively correlated with speed, and positively with repair measures.

**Accuracy:** WCR positively correlated with articulation rate, and negatively with total repair. Errors per 100 words also positively correlated with total repair.

**Lexical complexity:** D positively correlated with speed measures, and negatively with lexical errors; TTR positively correlated with frequency and length of mid-clause pauses. One measure of lexical sophistication, i.e., K3 words, correlated with length of mid-clause pauses.

**Metadiscourse markers:** Both metadiscourse marker type and token positively correlated with speed measures and frequency of end-clause pauses. Interestingly, there is a negative correlation between metadiscourse marker token and length and frequency of mid-clause pausing.

# 6. QUALITATIVE RESULTS (Research question 3)

As specified in Section 4.1. the qualitative analysis was to answer RQ3, as well as obtaining further insights to the quantitative findings reported in Section 5.

> **RQ3. Are there any differences between B2 and C1 candidates in the way pauses are used?**

Following the procedures explained in Section 4.4, the three researchers agreed on the three main categories and eight sub-categories summarised in Table 3.

*Table 3: Summary of main and sub categories for the use of pauses*

| Main category | Sub-category | Explanation |
|---|---|---|
| **1. Pauses related to access and retrieval difficulty** | Lexical Structure | **1a.** Mid-clause pauses for lexical/structural search which was followed by more sophisticated language<br>**1b.** Mid-clause pauses for lexical/structural search which however results in erroneous utterances or in generic expressions<br>**1c.** Pauses in the middle of / after producing sophisticated language |
| | Memory | **1d.** Pauses to recall items from long-term memory |
| **2. Pauses related to reformulations** | Reformulating | **2a.** Mid/end-clause pauses occurring during / before reformulating ideas and utterances, and making self-corrections |
| | Rescuing | **2b.** Mid-clause pauses in the middle of ungrammatical structures in the attempt of restructuring sentences |
| **3. Pauses related to effective speech delivery** | Topic development | **3a.** Pauses before adding more information, examples and justifications |
| | Attracting listeners' attention | **3b.** Mid-clause pauses before making evaluative comments and before expressing feelings (especially after an intensifier) |
| | Topic shift | **3c.** End-clause pauses before topic shift |
| | Dispreference | **3d.** Turn-initial pauses before dispreferred responses |

In the sections that follows, we will present each of the main and sub categories in details, by exemplifying the salient types of utterance environment identified for each category.

## 6.1 Pauses related to access and retrieval difficulty

### 1a Mid-clause pauses for lexical/structural search which was followed by more sophisticated language

The first sub-category under this theme was that some mid-clause pauses were observed before sophisticated language. A few examples are presented below.

In Excerpt 1, the C1 candidate, C048, produced a combined (filled + silent) pause of 1.28 seconds at a mid-clause location (line 3), before uttering the word, '*fortify'*. *Fortify* is a low frequency word which appears at the K6 band of the BNC. Band K6 was the lowest frequency band that C048 used in his entire utterance, and compared to the other three K6 words he produced (i.e. *soccer, stadium* and *Ukraine*), *fortify* seems a much more abstract lexical item.

Similarly, although pauses were shorter than 1 second, C048 also produced a 0.7-second mid-clause pause before '*consumed'* (line 2), which was one of the few words annotated as an Academic Word in his speech, as well as a relatively short 0.25-second pause before '*substances'*, a K3 word (line 2).

**[Excerpt 1] Candidate: C048, Level: C1**

1 **C048:** because sometimes they (0.29mu) they're (0.46mu) getting so excited **I** or even (0.72ec)
2 they have (0.7mv) consumed some (0.25mv) substances before the (0.67mc) event **I** or
3 (0.35ev) they can (0.64mv) burn (0.47mv) some materials **::** to to just (1.28mc) fortify this
4 emotions **::** to just like make them (0.3mu) stronger **I** and that is not very good **I**

Similarly, the lowest frequency word that C045 produced was '*tiresome'*, which is a K10–K20 word according to the BNC, and a mid-clause silent pause of 0.58 seconds was observed as in Excerpt 2.

**[Excerpt 2] Candidate: C045, Level: C1**

1 **C045:** the bus ride was long (0.3eu) **|** it was (0.58mu) tiresome

Among other examples, another C1 candidate (C039), as illustrated in Excerpt 3, paused before expressing conceptually more demanding ideas, when responding to one of the question points on why people dress in different ways. She had a 2.23-second and a 1.16-second mid-clause pause respectively, before and while talking about climate conditions (lines 1–2), and she paused before and while giving another example. She also had two end-clause pauses (1.83 seconds and 1.45 seconds) and one mid-clause pause (1.34 seconds) in lines 4–5. Before reformulating '*asian'* into '*asia'*, she also had a 1.25 end-close pause, but this is to be revisited when Category 2a is detailed below.

**[Excerpt 3] Candidate: C039, Level: C1**

1 **C039:** i think (2.74ec) **::** this is because (0.41ev) **::** we are living in (2.23mc) various (0.37mv)
2 climate (0.25mv) conditions (1.16mc) in the (0.37mv) asian (1.25ev) in asia **|** for instance
3 it's maybe very hot (0.73ec) **|** so (0.59ev) they would of course prefer (0.56mv) a light
4 clothing (0.52eu) **|** or (1.83ec) a st- arab (1.34mc) on the & in this picture (0.28eu) he is
5 using (0.37mv) a kaftan (1.45ec) because of the heat in his country (0.48eu) **|** and also
6 people in the north have **|**

Such use of pauses before sophisticated language or before/while expressing conceptually demanding ideas were more saliently observed among C1 students, although a couple of less obvious examples were also obtained among B2 students.

## 1b Mid-clause pauses for lexical/structural search which however results in erroneous utterances or in generic expressions

While those pauses contributing to the production of high-level language seemed to be a characteristic of C1 speech, it seems that similar attempts were also made by B2 learners. However, those long pauses to search lower frequency words tended to lead to unsuccessful outcomes among B2 learners.

Excerpt 4 presents a middle part of B029's responses that described her visit to a cathedral in Morelia, Mexico. In lines 3–4, it appears that she attempted to express how rare it is to visit such a magnificent, ancient cathedral. However, as in line 3, after a 1.04-second pause probably to look for an appropriate, specific phrase to continue, the search seemed to be unsuccessful. As a result, the utterance ended with the very generic word choice, '*a great places',* with an agreement error. Similarly, in line 5, a mid-clause pause of 1.66 seconds is observed after *'which is a very'.* Nevertheless, it seems that the candidate failed to come up with an appropriate word to elaborate on her perception of the building and abandoned to complete the relative clause.

**[Excerpt 4] Candidate: B029, Level: B2**

1 **B029:** and that building is famous for its (0.3mv) majestic (0.52mu) **|** there are (0.53mv)
2 building and (0.56mv) lamps in that place (0.48eu) **|** and it's really magical **::** to be there **|**
3 not many people have the opportunity the chance **::** to be (0.4mu) at such (1.04mc) a
4 great places (0.6eu) ……. i think **::** that (0.53ev) that building has probably (1.04ec) most
5 of the (0.53mu) history of aurelia Mexican in it (1.15ec) **::** which is a very (1.66mc) **|** it it
6 really hides a lot of …

Similarly, as a part of B034's talk on his last experience of visiting a tall building (see Excerpt 5), it appears that he wished to say that office buildings would need to be tall in order to accommodate all employees in limited land spaces for those buildings. However, it seems that after two long pauses in lines 2–3 (1.08 seconds and 1.48 seconds) to search for an appropriate verb (such as 'accommodate'), he gave up the search and ended up with producing an erroneous expression.

**[Excerpt 5] Candidate: B034, Level: B2**

1 **B034:** there are a lot of workplaces (0.35ev) **|** and they cannot be all fitted on (0.38mu) one tw-
2 on two or three (0.33mu) floors **|** so they need around twenty thirty floors **::** to (1.08mc)
3 to (1.48mc) to on a place (0.5mv) all (0.67mu) all the workers **|** but also they don't take a
4 large spaces **::** like if (0.36eu) they would if they had only (0.39mu) two or three floors

A number of similar examples were identified among other B2 transcripts. In Excerpt 6, B030 describes her amazing travelling experience in Europe. When she paused for 1.96 seconds after *many* in line 1, it seems that she was trying to buy time to search what to continue to express her encounters. Although it is not clear from the transcript whether she was attempting to search more specific vocabulary than *'people'*, the resultant utterance is a long mid-clause pause followed by a generic word.

**[Excerpt 6] Candidate: B030, Level: B2**

1 **B030:** i (0.42ec) i saw (0.4mu) so many cultures so many (1.96mc) people **::** in a in a way
2 that you can't found in one place (0.37eu) **|** and that is really really cool

As such, long pauses that seemed to be associated with less successful lexical/structural search and planning seemed to be typical of B2 speakers.

## 1c Pauses in the middle of / after producing sophisticated language

Related to the themes of the sub-categories 1a and 1b, it was interesting to observe that some long pauses occurred during or after producing sophisticated language.

For example, in Excerpt 7, the C1 candidate who was also exemplified in Excerpt 1 (C048) paused for 1.42 seconds between '*conflicting'* and '*behaviour'*. Although the two vocabulary items are not of very low frequency, placed in the BNC K2 band, *conflicting* is also classified in the Academic Word List. Perhaps more importantly, finding the right collocation, i.e. *conflicting behaviour*, seems much more challenging than producing these two words individually.

**[Excerpt 7] Candidate: C048, Level: C1**

1 **C048:** and that leads to their (0.45mu) to anger **::** and to conflicting (1.42mc) behaviour **I** so
2 people (0.32mu) can beat each other (0.53eu) on

On the other hand, B034 in Excerpt 8 paused for 1.6 seconds, after having managed to pronounce the low frequency word, *skyscrapers*, after some hesitations. The vocabulary is a K12 word, which was the lowest frequency item that B034 produced in his talk, and it sounds as if he needed to take a rest after the peak of his achievement!

**[Excerpt 8] Candidate: B034, Level: B2**

1 **B034:** and in those skp scap- scrapers (1.6mc) there are a lot of workplaces

### 1d Pauses to recall long-term memory

Here, it is important to note that the pauses classified in the above three sub-categories (1a, 1b and 1c) should not be confused with those pauses that were necessary to recall one's long-term memory. It is not always straightforward to distinguish pauses for lexical or structural search and pauses used for recalling memories without having stimulated retrospective interviews. However, when scrutinising examples of pauses, the three researchers had a clear consensus that the pauses categorised in the above three sub-categories looked very different from the ones for recalling memory as observed in **Excerpt 9, line 2 (i.e. trying to remember which countries were involved in the football match).**

**[Excerpt 9] Candidate: C040, Level: C1**

1 **C040:** however (0.43eu) the last (0.5mv) game i watched or briefly watched (0.42eu) **::** was
2 (1.14mv) the football (0.51mu) game between (0.76mc) Italy and Belgium last Monday

## 6.2. Pauses related to reformulations

The second main category of pauses associated with ways in which ideas and utterances are reformulated, and with the attempts to restructure ungrammatical structures. The former tended to be observed more among C1 speakers, while the latter was a typical characteristic of B2 speech.

### 2a Mid/end-clause pauses occurring during / before reformulating ideas and utterances, and making self-corrections

Excerpt 10 shows C038's description of how a teacher trainer gave feedback on C038's exceptional interpersonal ability. It seems that when C038 was quoting the feedback in line 4, she remembered the exact wording that the trainer used, and started reformulating the previous utterance. When doing so, she repeated '*no'* twice, with a pause of 1.16 seconds in between (line 4).

**[Excerpt 10] Candidate: C038, Level: C1**

1 **C038:** she observed **::** that i had a very (0.91mc) great ability **::** to understand people (0.35ev) **I**
2 she said **::** I was highly understanding (0.65eu) **I** and she said **::** that i was able i had that
3 power I had ability **::** to bring out people from their (0.29mu) souls **::** when they tense **I**
4 no (1.16ev) no i could soothe them (0.48eu) **I** i could bring them out of their problem **I**
5 give them motivation **I** give them courage

Another example is illustrated in Excerpt 11, where C044 described how his family had got lost in one of the Canary Islands. During his explanation of the challenges resulting from the identical or similar names given to different streets, '*every'* was self-corrected to be '*a lot of'* after a pause of 0.26 seconds (line 1), and *'the same names'* was reformulated into *'almost the same'* (line 2) in order to refine the accuracy of the description.

**[Excerpt 11] Candidate: C044, Level: C1**

1 **C044:** the main problem in thi- on this island is that **::** every (0.26ev) a lot of streets they have
2 the same names the same names (0.67eu) almost the same **::** you know
3 …
4 so he took us (0.26mu) to the place**I** first of all he took us in the wrong place **::** because it
5 had this street had the same name (0.61eu) **::** as we wanted (1.11ec) which we wanted **I**

While the above example might also be considered as the reformulation of ideas, a long end-clause pause of 1.11 seconds in line 5 seemed to play a slightly different role. When further elaborating on the experience of getting lost due to confusing street names, he exemplified how a taxi driver took his family to a different street that had the same name as the street they intended to reach. It seems that not only did C044 have a bit of hesitation pause of 0.61 seconds before forming the subordinate clause *'as we wanted'*, but he produced a long pause of 1.11 seconds in line 5 to rethink the structure that he used and corrected *'as'* into *'which'*. Unfortunately, the second syntactic structure is still problematic, but this pause of 1.11 seconds is very likely to be used by C044 to monitor his previous output and then decide to reformulate the previous utterance for enhancing accuracy. This type of reformulation is related to the next sub-category 2b, in which speakers pause to restructure an ungrammatical sentence.

## 2b Mid-clause pauses in the middle of ungrammatical structures in the attempt of restructuring sentences

In speech of B2 candidates, some long pauses of over 1 second were observed when they attempted to restructure an ungrammatical structure as in Excerpt 12, line 2. The difference between this and Excerpt 11 earlier is that pauses in this sub-category seemed to occur after an error was detected in the middle of an ungrammatical sentence.

**[Excerpt 12] Candidate: C030, Level: B2**

1 **B030:** and i went to a lot of (0.26mu) old buildings (0.39eu) to a lot of places **::** that told you
2 history (0.86ec)**I** i was a lot of cities **::** that (1.21mc) didn't end **::** because are so big **I**

While those pauses seemed to be used to buy some time in the attempt of rebuilding a correct syntactic structure, it was often the case that such long pauses also indicated that the grammatical structure was too wrong to be reformulated easily.

Instead, some pauses of this purpose were successfully utilised if the speaker had realised the complexity of syntactic parsing before an error was made. In Excerpt 13, it seems that B032 realised that she produced a very long subject in lines 1–2 and paused for 1.13 seconds after the long noun phrase. This appeared to help the speaker notice the need of a verb and the structure of the rest of the utterance was not distorted.

**[Excerpt 13] Candidate: B032, Level: B2**

1 **B032:** and any kind of (0.29mv) people around the world (0.31eu) for (0.63ev) millions of and
2 thousands of people (1.13mc) came around the world **::** to watch the world championship

## 6.3 Pauses related to effective speech delivery

Unlike the above two main categories, the final main category presents pauses that seemed to contribute to the effective delivery of speech.

### 3a Pauses before adding more information, examples and justifications

It was often the case that a long pause was observed before previous talk was elaborated with additional information to refine description and to broaden the scope of the talk.

Excerpt 14 exemplifies how C044 paused as he elaborated on the type of accommodation little by little (line 2), and before adding a justification (lines 4–5).

**[Excerpt 14] Candidate: B044, Level: C1**

1 **C044:** and we were trying **::** to find this flat by ourself (0.38eu) **::** cos it was booked (1.13mv)
2 from Airbnb (1.26ec) **I** and it was not the hotel (0.33eu) **I** a normal flat a- apartment
3 **…**
4 and we were students **I** we didn't want **::** to pay so much (1.27ec) **::** because it was like
5 low budget trip

Most of these pauses were found at end-clause locations or before adverbial phrases, which indicates that these pauses were likely to be used for searching and forming ideas rather than searching language, and these pauses sounded very natural.

However, when they were in a mid-clause position (see Excerpt 15, line 1), a skilful use of *'like'* was sometimes observed as in Excerpt 15, so that the lexical item can communicate to the listener that the speaker is looking for words to continue.

**[Excerpt 15] Candidate: C048, Level: C1**

1 **C048:** and (0.51ev) this (0.35mv) transfers from one person to others **I** they're (0.55eu) like
2 (1.07mv) **::** keeping each other by the hand and jumping

## 3b Mid-clause pauses before making evaluative comments and before expressing feelings (especially after an intensifier)

The data analysis suggested that some mid-clause pauses were effectively utilised to have a certain communicative impact on listeners. This usage of long pauses was observed in both B2 and C1 speech. For instance, Excerpt 16 shows how C048 described the excited audience of a football match. In C048's entire speech production for this task, he used 10 mid-clause pauses over 1 second in total, but 4 of them were clustered in the two lines presented in Excerpt 16, where he highlighted the emotional state of the crowd. It should be noted that the intensifier, *'very'*, is repeated twice before the full clause '*it is very intense'* was uttered.

**[Excerpt 16] Candidate: C048, Level: C1**

1 **C048:** so (0.49eu) it is (0.65mv) very (1.01mv) very (1.68mc) it is very intense (2.41ec) by
2 mentally and (0.74mv) psy- (1.23mv) by (2.0mc) by the feelings **I** it is very intense

Also, a number of examples were observed where candidates paused at mid-clause locations before uttering emotionally charged words. For example, in Excerpt 17, where B032 described her experience in watching the football World Cup, and she inserted a pause of 0.94 seconds and 0.41 seconds respectively before uttering *'amazing'* and *'enthusiastic'* (line 1). A pause of 0.43 seconds was also observed before *'exciting'* in line 2. Here it is notable that she had pauses between an intensifier and an adjective, i.e. between *'so'* and *'amazing'* and between *'very'* and *'exciting'*. It seems that this was a common phenomenon, which will be further exemplified below.

**[Excerpt 17] Candidate: B032, Level: B2**

1 **B032:** it was so actual- it was so (0.94mc) amazing (0.41mu) and enthusiastic … that was a very
2 (0.43mc) exciting moment ever for me

In Excerpt 18, when B030 was describing her experience of travelling in Europe, and how astonishing the Pisa tower was, she described it as *'really amazing'* in line 1 but it seems that she wanted to revisit the phrase to further emphasise how impressive it was. Before uttering *'awesome'* in line 2, she effectively paused for 0.68 seconds.

**[Excerpt 18] Candidate: B030, Level: B2**

1 **B030:** and i will told you about the pisa tower (0.9eu) **I** it was really really amazing**I** i don't know i
2 i (0.32eu) don't know **::** in what way describe it **::** because it was (0.68mu) awesome

Such evaluative adjectives often appeared after a mid-clause pause. In Excerpt 19, a pause of 0.34 seconds appeared between *'pretty'* and *'good'*.

**[Excerpt 19] Candidate: B031, Level: B2**

1 **B031:** and (0.68eu) yeah it was it was football (0.4eu) **::** like the champion cup is at the
2 moment in France (0.59eu) **I** they played in paris **I** it was it was (0.59mu) pretty (0.34mu)
3 good

Here is another example of the same kind. Excerpt 20 shows B035's talk that described an injured football player. B035's talk for this task was characterised as no occurrence of long mid-clause pauses over 1 second, and the longest mid-clause pause he had was 0.53 seconds observed in line 2. This longest mid-clause pause was used when he offered his view to the incident, uttering *'too (0.53mc) too serious'.* This sounded very effective to express his evaluative summary of the event that he had talked thus far.

**[Excerpt 20] Candidate: B035, Level: B2**

1 **B035:** and as i as i heard from the news **::** he was almost killed in this (0.33mu) encounter **|** so
2 (0.31eu) i think **::** that some people take sports too (0.53mc) too serious (0.36eu)

## 3c End-clause pauses before topic shift

It seems that C1 speakers were particularly good at using pauses at end-clause locations to signal topic shift. For instance, Excerpt 21 clearly shows that a long end-clause pause of 1.62 seconds was placed before moving on to the next topic in the prompt, *'what did you feel about watching this event'*, which C040 read aloud from the screen.

**[Excerpt 21] Candidate: C040, Level: C1**

1 **C040:** especially when (0.42mv) belgium is is (0.64mv) playing (0.57eu) **|** so (1.62ec) what did i
2 feel about **::** watching this event (1.28ev) **|** actually i i felt very bored (0.42ev) **::** because
3 i'm not really into it (0.43eu)

Excerpt 22 is a similar example, where C039 inserted a long pause of 2.94 seconds before initiating a new topic by reading aloud the next prompt, *'why do people dress in such different ways?'.*

**[Excerpt 22] Candidate: C039, Level: C1**

1 **C039:** very common (0.57mv) these days **::** (2.94ec) why people dress in such different ways **|**
2 i think (2.74ec) **::** this is because (0.41ev) **::** we are living in…

Equally, end-clause pauses were also used to signal a shift of scope in related topics. For example, in Excerpt 23, till the end of line 2, C045 was talking about his travel experience with a friend. But after the long end-clause pause of 2.1 seconds at the end of line 2, he moved from personal experience to a general description of how people travel in the country.

**[Excerpt 23] Candidate: C045, Level: C1**

1 **C045:** looking at people **::** sitting by the river (0.72eu) **::** doing (0.31eu) going about their daily
2 chores (0.27eu) **::** it was a fascinating experience for all of us (2.1eu) **|** people in this
3 country usually travel long distances by bus and (0.26mu) train (0.75eu)

**Excerpt 24** is the final example of using a long end-clause pause to signal topic shift. C047 here effectively used a long pause of 1.26 seconds (line 1) before making a concluding remark by noting that he finished responding to all three required elements of the long turn task.

**[Excerpt 24] Candidate: C047, Level: C1**

1 **C047:** and more likely you will need it at some point in time (1.26eu) **|** so (0.35eu) this is
2 (0.51mv) what i can say (0.31mu) about these three questions **|** so what do i do **|**

This observation is in line with the CA literature, which the use of long pauses, usually exceeding one second, has long been discussed. Examining a wide range of two- to multi-party interactions, Jefferson (1989: 192–193) concluded that the ['standard maximum' silence of approximately one second] can serve a certain role in a conversational sequence and one of such roles is to signal topic shift.

## 3d Turn-initial pauses before dispreferred responses

The last sub-category of this main theme refers to a long pause at a distinct position – the turn initial position of the response time, as shown in Excerpt 25. In this excerpt, C040 was requested to describe her experience in watching a sport match. She hesitated at the turn initial part with a pause of 1.04 seconds, as in *'I (1.04) I'm not really into actually sports at all…'*.

**[Excerpt 25] Candidate: C040, Level: C1**

1 **C040:** i (1.04mv) i'm not really into actually sports at all (0.5eu) **I** and i don't watch that very
2 often on tv

This quote demonstrates that the notion of *preference* in Conversation Analysis (CA) is applicable even in a monologic test task context. The preference structure in CA characterises conversational properties when alternative types of actions are available to the conversants, but these alternatives are non-equivalent (Atkinson & Heritage, 1984: 53). For instance, 'offers' can be accepted or refused, 'assessments' can be agreed or disagreed with, and 'requests' can be granted or declined. The structural characteristics of talk-in-interaction are designed to prefer one of the actions (e.g. acceptance, agreement or granting), and to disprefer the other alternative actions. It should be noted that the notion of preference here is not intended to refer to the psychological dispositions or motives that the participants may personally prefer, but to structural characteristics of talk-in-interaction designed for particular actions (Hutchby & Wooffitt, 1998: 43–44). For instance, if there is a preference for acceptance after an invitation, it is an institutionalised preference bearing on that choice itself and not a characterisation of participants' desires.

Dispreferred actions are typically delayed at the beginning of the utterance, using filled or silent pauses or other hesitation markers. Excerpt 25 exemplifies that her refusal to the task 'request' was indeed dispreferred. It is interesting that her dispreferred response even under a computer-delivered monologic test condition was delivered with hesitation features at the turn initial position of her talk, as we would normally do in real-life conversation.

# 7. DISCUSSION AND CONCLUSIONS

The primary aim of the current study was to help develop a better understanding of oral fluency and its relationship to other aspects of performance at B2 and C1 level in the Aptis Speaking test. Against the backdrop of the findings of Tavakoli, et al. (2017) reporting non-significant differences between different aspects of fluency in B2 and C1 level performances, this study set out to explore what aspects of performance (i.e. lexical complexity, syntactic complexity, accuracy or metadiscourse marker features) are key in distinguishing B2 and C1 levels of proficiency in the Aptis Speaking test. In what follows we will summarise the findings of the study and discuss them in the light of the literature reported earlier on in the paper. It is necessary to note that given the small sample size of the study, results should be interpreted with care.

## 7.1 Important role of accuracy and lexical diversity (RQ1)

The first research question was concerned with the linguistic characteristics of performance in terms of lexical and syntactic complexity and accuracy that distinguished the two levels of B2 and C1. The results of the *t*-tests indicated that measures of accuracy (WCR and number of lexical errors) and lexical diversity (TTR and D) discriminated candidates between B2 and C1. For all the significant results, large effect sizes of above 1 were observed. The differences between the two groups for syntactic complexity, however, did not reach significant results. These findings suggest that at higher levels of proficiency, although measures of utterance fluency do not seem to distinguish the two proficiency levels, measures of accuracy and lexical diversity play an important role in discriminating the levels. This is an important finding as it indicates that while B2 and C1 levels are different for some aspects of their performance, i.e. accuracy and lexical diversity, they are not different in terms of fluency and syntactic complexity of their language output. This research therefore validated part of the current Task 4 rating descriptors that refer to these aspects (see below):

> **C1:** 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.
>
> **B2.2:** 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.
>
> O'Sullivan and Dunlea (2015: 63)

The findings of Tavakoli, et al. (2017) indicating a ceiling effect for fluency across higher levels of proficiency is more recently replicated by Tavakoli, Slaght, Kendon and Hunter (forthcoming) examining fluency in Test of English for Educational Purposes (TEEP) speaking test. As regards syntactic complexity, the findings of the current study may suggest that there is a similar ceiling effect for syntactic complexity in higher levels of proficiency. Alternatively, it is possible to hypothesise that at higher levels of proficiency, we may need more fine-tuned measures of syntactic complexity to distinguish proficiency levels. Norris and Ortega (2009) have argued that subordination is a reliable measure of complexity especially for intermediate levels, but for advanced levels of proficiency, other measures such as the degree of the phrasal complexity might be a more appropriate measure of syntactic complexity. Therefore, further research with different syntactic complexity measures would be useful to (dis)confirm that the non-significant results obtained for syntactic complexity in this study is not an artefact of the choice of the measures.

## 7.2 Correlations with fluency measures (RQ2)

Our second research question asked whether the linguistic measures of syntactic and lexical complexity and accuracy and use of discourse markers correlated with various fluency measures from Tavakoli et al. (2017). The summary of the findings of the correlations suggests:

- longer clauses were associated with shorter stretches of connected speech, slower speed and more repairs
- longer AS-units were linked with more end-clause pauses
- more accurate language contained fewer repairs and was faster; more accurate language was in the form of shorter clauses with fewer lexical errors
- more errors were associated with more repairs and slower speed
- lexically more diverse language was faster, but it contained more lexical errors
- lexically more diverse language correlated with longer and more frequent mid-clause pauses
- lexically more diverse language contained more mid-clause filled pauses
- more varied use of metadiscourse markers was associated with faster speech, more frequent end-clause silent pauses and fewer repetitions.

The results of correlations suggested a few typical trends of speech production also reported by previous research. For example, speed fluency negatively correlates with length of clause (Tavakoli, 2018); longer end-clause pauses correlate with longer AS-units (Tavakoli, 2011); and more repair is seen in longer clauses (Awwad & Tavakoli, 2019). The finding that C1 level compared to B2 level paused more at end-clause position is in line with previous research (Kahng, 2014; Tavakoli, 2011), suggesting that more proficient speakers (as well as native speakers) pause more frequently at end-clause rather than mid-clause positions. The negative correlation between speed fluency and length of clause is also interesting as it implies that L2 speakers' attempt at producing longer clauses might have a damaging effect on their speed. In other words, producing shorter clauses seems to provide an opportunity for a faster and more accurate performance.

As the relationship between fluency and accuracy is concerned, WCR positively correlates with articulation rate, and negatively with repair, suggesting that speakers with higher speed regularly produce more accurate clauses. Number of lexical errors in 100 words is also correlated positively with repair, and negatively with articulation rate, implying that lexical errors invite more repair and make the speech slow. This finding is linked with the assumptions of the monitoring processes in speech production in general (Levelt, 1989) and with the principles of error-treatment (Ahmadian, Abdolrezapour & Ketabi, 2012; Kormos & Sáfár, 2008) in particular (see discussion below). Measures of lexical diversity positively correlate with speed and breakdown measures, and negatively with lexical errors. These findings suggest that those speakers with a higher level of lexical diversity are faster in their speech, but make more mid-clause pauses and fewer repairs. Previous research has shown that lexical knowledge is a reliable predictor of proficiency (Daller & Xue, 2007; Revesz et al., 2016), and that more proficient speakers have faster access to and retrieval of lexical items (Kormos, 2006). As the correlations between fluency and metadiscourse markers indicate, metadiscourse type positively correlates with both speed fluency and end-clause pauses, but negatively with mid-clause pauses and repair. Similarly, metadiscourse token correlates positively with speed and end-clause pauses, but negatively with mid-clause pausing. This suggests that efficient use of metadiscourse markers is linked with speedy performance, i.e. performance that is not interrupted by mid-clause pauses but supported by the end-clause pausing opportunity. This is a novel and interesting finding as it suggests using more metadiscourse markers (both type and token) may facilitate the speech production process as it is associated with more speed and less mid-clause pausing and repair.

## 7.3 Differences in the use of pauses between B2 and C1 (RQ3)

Our final research question asked whether there were any differences between B2 and C1 candidates in the way they paused during task performance. The results of the qualitative analysis provided interesting results. They suggested that pauses were related to access and retrieval difficulty the speakers experienced when performing the tasks. They often paused in search of lexical and structural items, and when attempting use of more sophisticated language. Such attempts, whether successful or less successful, were associated with mid-clause pauses highlighting the link between mid-clause pauses and Levelt's (1989) Formulation stage. Long pauses in this study were also observed when the speakers were trying to retrieve idea units from the long-term memory.

There is strong research evidence in SLA to suggest that for L2 learners, whose lexical, syntactic and phonological knowledge is still developing, their access to L2 knowledge is not yet automatic (Kormos, 2006). This less-than-automatic knowledge makes speech slower when disfluencies such as pauses and repairs emerge. Our results confirm the findings of this body of research as our qualitative analysis suggested that many of the pauses were linked with the need to monitor output (e.g., error detection and self-correction) and an attempt to repair it.

Research in SLA strongly recommends that a distinction is made between mid-clause and end-clause pauses as they seem to be crucial in understanding the differences between L1 and L2 production. This body of research (Foster & Skehan, 1996; Tavakoli, 2011) has shown that L2 speakers pause more frequently at mid-clause position, while L1 speakers pause regularly at end-clause position. The mid-clause pauses are believed to be closely linked with the Formulation stage of the speech production process (Levelt, 1989) in which the pre-verbal message produced at the Conceptualisation stage is turned into language forms. L2 speakers are believed to pause in mid-clause positions to facilitate access and retrieval of linguistic items, to search for lexical and structural units that allow them to express their intended message, and to reformulate their language (Skehan, 2015; Skehan, Shum & Foster, 2019). In addition, there is emerging research evidence (de Jong, 2016) to suggest that both L1 and L2 speakers pause longer before a low-frequency lexical item than they do before a high frequency item. Our findings are in line with this body of research as we have several examples in the data set to show the L2 speakers pause longer before less frequent words (e.g. words from K6 and K12 lists).

The second important finding of the qualitative analysis is with regard to reformulations. Mid-clause pauses were occurring before and during reformulations, restructuring and self-corrections. Another important reason for mid-clause pauses in L2 speech is the activation of the monitoring processes (Kormos, 2006; Levelt, 1989). In L2 studies, monitoring, or more technically put 'self-monitoring', refers to L2 speaker's effort to check their speech in order to identify an error, an inappropriate aspect of their speech, or improve their utterances for better communication impact before or after language is produced (Levelt, 1983, 1999). The results of the qualitative analysis suggested that both B2 and C1 level speakers spent time monitoring their output typically reflected through mid-clause pauses and repairs. There are several examples of reformulation, our second category of the qualitative analysis, to suggest the speakers are engaged in the processes of error-detection and error treatment during the monitoring process. There were also several examples in the C1 level performances to suggest the speakers were involved in monitoring their own output for appropriacy and more effective communication reasons. This brings us to the final category of our findings that suggests pauses were used to produce more effective speech, i.e. to provide opportunities for topic development, to justify and evaluate points of discussion, to indicate topic shift, to intensify feelings, and to adhere to rules in conversation. This finding is particularly important as mid-clause pausing for some of these purposes is a characteristic of the C1 level's behaviour in this study.

Some key differences were observed between the way pauses were employed by B2 and C1 level speakers in the current study. As indicated in the results, a main reason for long mid-clause pauses occurring in the speaker speech was access to and retrieval of lexical items or syntactic structures. One difference noticed was that the C1 group seemed to be more successful in using the pausing opportunity to produce correct lexical items or syntactic structures, whereas the B2 level speakers' attempts proved to be less successful in this regard. Another difference observed between the two groups was their priority choices when using pauses. While the C1 level used the pausing opportunity to make their speech more effective, the B2 level mainly used pauses to search for linguistic items and to monitor their speech. Instead of spending time in correcting minor errors, the C1 level speakers used pauses to have more impact on the listener. In Excerpt 25, for example, the C1 speaker makes a long pause before saying *'i'm not really into actually sports at all'* to mark the discourse that her response is considered as a dispreferred choice in conversation. To sum up the differences between the two groups, it can be argued that while both B2 and C1 levels used pauses for access, retrieval and reformulation purposes, the category of pauses that make speech more effective is mainly used by C1 level speakers.

## 7.4 Additions to the C1 rating descriptor

Another aim of the project was to understand in what ways a careful discourse-analysis approach to analysing speaking performance at higher proficiency levels can help develop an in-depth understanding of criterial features of performance in the Aptis Speaking test. The findings of the study suggest while Tavakoli et al.'s (2017) quantitative study did not provide any useful fluency feature that could differentiate B2 and C1 candidates, the use of pauses is indeed different between the two groups. In doing so, this study demonstrated complex multi-dimensional nature of performance at the C1 level. That is, to differentiate C1 from B2, rating descriptors have to embrace and denote different language aspects even when the focus of an individual descriptor is one analytic aspect of the language. For example, a possible short descriptor that could be added to the C1 level of the Task 4 fluency rating scale based on the current research is:

- **C1:** Pauses are effectively used before successful production of sophisticated language or reformulation, and to make effective communicative effects on the listeners.

While this descriptor is most relevant to the Aptis Speaking test, it is likely that a similar descriptor can be appropriately used for other speaking tests that use monologic tasks whose target levels cover C1. This study also offered 25 excerpts to illustrate the similarities and differences between B2 and C1 speakers in terms of their use of pauses. It is believed that these excerpts will be useful as supplementary material for the Aptis rater training program, to raise the raters' awareness of how different language aspects might be interwoven, and how C1 speakers' proficiency is displayed through their multi-dimensional performance.

# REFERENCES


Ahmadian, M.J., Abdolrezapour, P., & Ketabi, S. (2012). Task difficulty and self-behaviour in second language oral production. *International Journal of Applied Linguistics*, 22, 310–330.

Atkinson, J.M., & Heritage, J. (1984). *Structures of social action.* Cambridge, New York: Cambridge University Press.

Awwad, A., & Tavakoli, P. (2019). Task complexity, language proficiency and working memory: Interaction effects on second language speech performance. *International Review of Applied Linguistics (IRAL).* DOI: https://doi.org/10.1515/iral-2018-0378

Bax, S. (2012). *Text Inspector.* Online text analysis tool. Available at https://textinspector.com/

Bax, S., Nakatsuhara, F. & Waller, D. (2019). Researching metadiscourse markers in candidates' writing at Cambridge FCE, CAE and CPE levels, *System, 83,* 79–95.

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

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, USA.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA* (Vol. 32, pp. 21–46). Amsterdam: John Benjamins.

Bygate, M., & Samuda, V. (2005). Integrative planning through the use of task repetition. In R. Ellis, *Planning and task performance in a second language* (pp. 37–74), Amsterdam: John Benjamins.

Cobb, T. (2013). *Web Vocabprofile*. Retrieved from http://www.lextutor.ca/vp

Cobb, T., & Free, P. (2015). *Compleat lexical tutor* v. 8 [computer program]. Retrieved from http://www.lextutor.ca.

Creswell, J. W., & Plano Clark, V. L. (2011) *Designing and conducting mixed methods research (2nd edition),* Thousand Oaks, CA: Sage.

Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge.* Cambridge: Cambridge University Press.

De Jong, N.H. (2016). Predicting pauses in L1 and L2 speech, *International Review of Applied Linguistics in Language Teaching, 54*(2), 113–132.

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

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., & Vercellotti, M. (2016). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research, 20*(3), 387–404.

Ellis, R., & Barkhuizen, G. P. (2005). *Analysing learner language*. Oxford University Press, USA.

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., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied linguistics*, *21*(3), 354–375.

Foster, P., & Wigglesworth, G. (2016) Capturing accuracy in second language performance: The case for a weighted clause ratio, *Annual Review of Applied Linguistics 36*, 98–116.

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

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

Hasselgreen, A. (2004). *Testing the Spoken Language of Young Norwegians.* Studies in Language Testing 20. Cambridge: UCLES/Cambridge University Press.

Have, P. ten (2006). Conversation Analysis Versus Other Approaches to Discourse. Review Essay: Robin Wooffitt (2005). Conversation Analysis and Discourse Analysis: A Comparative and Critical Introduction [32 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 7*(2), Art. 3, http://nbn-resolving.de/urn:nbn:de:0114-fqs060239.

Heritage, J. (1995). Conversation analysis: Methodological aspects. In U.M. Quasthoff (Ed.), *Aspects of Oral Communication* (pp. 391–418). Berlin: De Gruyter.

Heritage, J. (1997). Conversation analysis and institutional talk: Analysing data. In D. Silverman (Ed.), *Qualitative Research: Theory, Method and Practice* (pp. 161–182). London: Sage.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30 (4)*, 461–473.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA* (Vol. 32, pp. 1–20). Amsterdam: John Benjamins.

Hutchby, I., & Wooffitt, R. (1998). *Conversation Analysis.* Cambridge: Cambridge University Press.

Hyland, K. (2004). *Disciplinary Discourses: Social Interactions in Academic Writing.* Michigan: University of Michigan Press.

Inoue, C. (2016). A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research. *The Language Learning Journal, 44*(4), 487–505.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008) Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics 29*(1): 24–29.

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, British Council, London. 1–93.

Jefferson, G. (1989). Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In D. Roger & P. Bull (Eds.), *Conversation: an interdisciplinary perspective* (pp. 166–196). Clevedon: Multilingual Matters.

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

Koponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), Perspectives on fluency (pp. 5–24). Ann Arbor: University of Michigan Press.

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

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and cognition, 11*(2), 261–271.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests.* Studies in Language Testing 14. Cambridge: UCLES/Cambridge University Press.

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

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25–42). Ann Arbor: University of Michigan Press.

Levelt, W.J.M. (1983). Monitoring and self-repair in speech. *Cognition, 41,* 41–104.

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

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85–104.

McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal, 1,* 1–15.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20*(1), 83–108.

Michel, M. (2017). Complexity, accuracy, and fluency in L2 production. In *The Routledge Handbook of Instructed Second Language Acquisition* (pp. 66–84). New York: Routledge.

Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance, *Language Testing, 31*(2), 147–175.

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

Nakatsuhara, F. (2014). *A Research Report on the Development of the Test of English for Academic Purposes (TEAP) Speaking Test for Japanese University Entrants – Study 1 and Study 2*, available online at: www.eiken.or.jp/teap/group/pdf/teap_speaking_report1.pdf

Norris, J.M., & L. Ortega. (2009). Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Applied Linguistics 30*(4), 555–578.

O'Sullivan, B., & Dunlea, J. (2015). *Aptis General 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

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse processes*, 43(2), 121–152.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics 30*(4), 590–601.

Préfontaine, Y., Kormos, J., & Johnson, D. E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing*, *33*(1), 53–73.

Read, J. (2000). *Assessing Vocabulary.* Cambridge University Press.

Révész, A., Ekiert, M., & Torgersen, E. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, *37*(6), 828–848.

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of child language, 14*(02), 201-209.

Richards, K., & Seedhouse, P. (2005). *Applying Conversation Analysis.* Basingstoke: Palgrave Macmillan.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language 50*(4), 696–735.

Schegloff, E.A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction 26*(1), 99–128.

Schiffrin, D., Tannen, D., & Hamilton, H. (2001). *The Handbook of Discourse Analysis.* Oxford: Blackwell.

Seedhouse, P., & Nakatsuhara, F. (2018). *The Discourse of the IELTS Speaking Test: The Institutional Design of Spoken Interaction for Language Assessment*, Cambridge: Cambridge University Press.

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

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

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

Skehan, P. (2015). Limited Attention Capacity and Cognition. In M. Bygate (Ed.), *Domains and Directions in the Development of TBLT* Vol. 8 (pp. 123–155). Amsterdam: John Benjamins Publishing.

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

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 133–150.

Tavakoli, P. (2018). L2 Development in an intensive Study Abroad EAP context, *System, 72*, 62–74.

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., & Hunter, A.M. (2018). Is fluency being 'neglected' in the classroom? Teacher understanding of fluency and related classroom practices. *Language Teaching Research, 22*(3), 330–349.

Tavakoli, P., Nakatsuhara, F., & Hunter, A.M. (2017). *Scoring validity of the Aptis Speaking test: Investigating fluency across tasks and levels of proficiency*, ARAGs Research Reports Online, AR-G/2017/7, British Council, London, available online at: www.britishcouncil.org/sites/default/files/tavakoli_et_al_layout.pdf

Tavakoli, P., & Skehan, S. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (ed.) *Planning and Task Performance in a Second Language* (pp.239–273). John Benjamins.

Wooffitt, R. (2005). *Conversation Analysis and Discourse Analysis: A Comparative and Critical Introduction.* London: SAGE Publications.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and online planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics, 24*(1), 1–27.

Yin, R. K. (2011). *Qualitative research from start to finish.* New York: Guildford Press.

# Appendix A:
## Tavakoli et al.'s (2017) suggested fluency descriptors

*Table A-1: Tavakoli et al.'s (2017) suggested fluency descriptors for Task 1*

| | | |
|---|---|---|
| **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 A-2: Tavakoli et al.'s (2017) suggested fluency descriptors for Tasks 2 and 3*

| | | |
|---|---|---|
| **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 A-3: Tavakoli et al.'s (2017) suggested fluency descriptors for Task 4*

| | | |
|---|---|---|
| **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). |

# Appendix B: Coding symbols

The following symbols were used to annotate transcripts in this study.

| | |
|---|---|
| \| | AS-unit boundary |
| **::** | Clause boundary |
| **≠** | Global error |
| **≠~** | Lexical error |
| **errfr** | Error-free clause |
| **(.8)** | Level **1** error |
| **(.5)** | Level **2** error |
| **(.1)** | Level **3** error |
| **$** | End clause filled pause |
| **$*** | Mid clause filled pause |
| **^** | Repetition |
| **&** | Self repair |
| **%** | Reformulation |
| **@** | Laughter |
| **#** | Cough/clear throat |
| **!** | False start |
| **mc** | Mid-clause combined (filled + silent) pause |
| **ec** | End-clause combined (filled + silent) pause |
| **mv** | Mid-clause filled pause |
| **ev** | End-clause filled pause |
| **mu** | Mid-clause silent pause |
| **eu** | End-clause silent pause |

## British Council Assessment Research Group

The Assessment Research Group was formed in 2013 to support the British Council's work in assessment and testing across the world. The team is responsible for ensuring that all new assessment products and new uses of existing products are supported by the most up-to-date research. They also continuously evaluate the quality of British Council assessment products.

# TOWARDS A MODEL OF MULTI-DIMENSIONAL PERFORMANCE OF C1 LEVEL SPEAKERS ASSESSED IN THE APTIS SPEAKING TEST

VS/2019/001

**F. Nakatsuhara, P. Tavakoli and A. Awwad**

**www.britishcouncil.org/aptis/research**

BRITISH COUNCIL VALIDATION SERIES

Published by British Council
10 Spring Gardens
London SW1A 2BN

**© British Council 2019**
The British Council is the United Kingdom's international organisation for cultural relations and educational opportunities.