5-4-2021

The Relationships between Second Language Speakers’ Oral Productions, Oral Proficiency, and their Individual Differences: A Longitudinal Study

Tamanna Mostafa
Georgia State University

Recommended Citation

Mostafa, Tamanna, "The Relationships between Second Language Speakers’ Oral Productions, Oral Proficiency, and their Individual Differences: A Longitudinal Study." Dissertation, Georgia State University, 2021.
doi: https://doi.org/10.57709/22259031

THE RELATIONSHIPS BETWEEN SECOND LANGUAGE SPEAKERS' ORAL PRODUCTIONS, ORAL PROFICIENCY, AND THEIR INDIVIDUAL DIFFERENCES: A LONGITUDINAL STUDY

by

TAMANNA MOSTAFA

Under the Direction of YouJin Kim, PhD

ABSTRACT

Despite the importance of English speaking skills in higher education contexts (Andrade 2009), there has been a lack of investigations into longitudinal development in English as second language (ESL) speakers' oral proficiency in relation to their oral production features (complexity, accuracy, fluency: CAF) and individual differences in working memory (WM) and aptitude. Existing research examining the relationships between CAF measures and L2 oral proficiency mostly focused on monologic tasks although CAF measures might significantly vary between monologic and dialogic task types (Michel et al., 2012). The purpose of this dissertation is threefold. First, the study investigates whether CAF measures of ESL speakers' monologic and dialogic oral performances predict development in their oral proficiency over time. Second, the dissertation examines whether ESL speakers' WM and aptitude are predictive of their oral proficiency development. Third, the dissertation also examines whether the relationships between CAF measures and oral proficiency are mediated by the speakers' WM and aptitude. In total, 60 ESL participants (matriculated and non-matriculated) performed both monologic and dialogic oral tasks at three different times over eight months. The participants' oral proficiency was measured by TOEFL iBT speaking tests and communicative adequacy ratings of their monologic and dialogic speech. The results show that in monologic speech, high proficient ESL speakers produced more syntactically and lexically complex language, whereas in dialogic speech, they produced faster speech. The findings also indicate that although in both monologic and dialogic speech, the participants with lower phonation (compared to pauses) significantly developed their oral proficiency over time, in dialogic speech, the participants with longer turns (in-between pauses) had longitudinal development in oral proficiency. The dissertation also found that high proficient ESL speakers with higher aptitude used more familiar vocabulary in their monologic speech but shorter fluent runs and shorter clauses in dialogic speech. Overall, the study argues that high proficient speech in monologic versus dialogic modes have different linguistic benchmarks. The findings also offer insights into the processes of high proficient L2 speech production in monologic and dialogic tasks by suggesting the combined effects of ESL speakers' aptitude and CAF features on their oral proficiency scores.

INDEX WORDS: Longitudinal development in oral proficiency, Complexity, Fluency, Aptitude, English as second language

THE RELATIONSHIPS BETWEEN SECOND LANGUAGE SPEAKERS' ORAL PRODUCTIONS, ORAL PROFICIENCY, AND THEIR INDIVIDUAL DIFFERENCES: A LONGITUDINAL STUDY

by

TAMANNA MOSTAFA

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2021

THE RELATIONSHIPS BETWEEN SECOND LANGUAGE SPEAKERS' ORAL PRODUCTIONS, ORAL PROFICIENCY, AND THEIR INDIVIDUAL DIFFERENCES: A LONGITUDINAL STUDY

by

TAMANNA MOSTAFA

Committee Chair: YouJin Kim

Committee: Scott Crossley
Eric Friginal
Andrea Révész

Electronic Version Approved:

Office of Graduate Services
College of Arts and Sciences
Georgia State University
May 2021

DEDICATION

To my little baby girl, Unaysa Hyder, my patient companion in my PhD journey. Hope you'll feel proud of your mama one day

ACKNOWLEDGEMENTS

First of all, I want to acknowledge the funding sources that helped me conduct the dissertation project: Language Learning Dissertation Grant, ETS Small Grants, and Provost dissertation Fellowship.

I want to express my heartfelt gratitude to everyone who helped me reach the end of my PhD journey. I want to thank my advisor Dr. YouJin Kim for her guidance throughout the process. If I compare my present self with the one who started the PhD program in Fall 2016, I see two totally different persons: my present self is way stronger in terms of intellectual maturity and the ability to combat obstacles and problems. Dr. Kim was one of those people who helped me achieve this transformation and establish myself as a researcher in the field. I acknowledge all the support I received from Dr. Kim not only in the process of publishing multiple papers in peer-reviewed journals but also in the process of conducting my dissertation project. Thank you, Dr. Kim for all your guidance and support, for understanding me, and for understanding my struggles. Next, I want to thank Dr. Scott Crossley for all the support I received from him in reaching my current position. I worked with Scott in several projects, and each of those was a huge learning experience for me. It was in Scott's quantitative research methods class that I found my knack for quantitative analysis, and that self-discovery (better late than never) helped me make significant career choices. I want to thank Dr. Eric Friginal for always being so encouraging and positive. I will remember Eric as one of the most supportive and encouraging professors I have been in touch with. I also want to thank Dr. Andrea Révész for graciously agreeing to be in my dissertation committee and extending her helping hand to me whenever I needed her support. In addition to my committee members, I want to thank all the faculty members in the Applied linguistics and ESL department at Georgia State University for helping

me grow as a scholar. I thank Dr. Diane Belcher for introducing me to the world of qualitative research, Dr. Ute Roemer for teaching me various techniques of corpus linguistics, Dr. Stephanie Lindemann for helping me be skilled in linguistic analysis, and Dr. Sara Cushing for teaching me advanced concepts in assessment. I also thank Dr. Viviana Cortes for kindly participating in preparing the Spanish version of a test for my dissertation study. In addition, I acknowledge the assistance and support I received from all my friends in the department throughout my PhD journey.

Last but not the least, I am grateful to my parents for instilling in me the drive to always do better. I also want to acknowledge the support I received from my dear siblings, friends, and cousins in Bangladesh who always believed in me. Finally, I would not have been able to finish the dissertation project without the mental support I received from my husband, Sayeed, and my daughter, Unaysa. Thanks, Sayeed for always being there for me. Thanks, my baby Unaysa for enduring everything I put you through. My achievement is yours.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Every year thousands of English as second language (ESL) speakers are enrolled in undergraduate or graduate degree programs in the USA, although many of them still struggle with meeting the demands of English in academic life (Andrade, 2006, 2009). Additionally, there are non-matriculated English language learners, taking classes in intensive English programs (IEP) in the US universities, who want to improve their English proficiency enough to get admission to undergraduate or graduate degree programs. There has been empirical evidence that ESL speakers in academic contexts face more challenges in developing English-speaking skills than any other skills (Andrade, 2006; Ferris, 1998; Ferris & Tagg, 1996a, 1996b). Lack of adequate English-speaking skills might negatively affect ESL speakers' academic and social adjustments in academic contexts (Andrade, 2009). Despite such importance of English-speaking skills in higher education contexts, there has been a lack of scholarly interests into investigating longitudinal development in second language (L2) oral proficiency (cf. Vercellotti, 2017; Tonkyn, 2012). ESL speakers' oral proficiency might be related not only to linguistic features of their oral production but also to their individual difference variables, for example, working memory (WM) and aptitude. In this dissertation, ESL speakers refer to both non-matriculated and matriculated L2 speakers of English. The dissertation study focuses on the constructs of ESL speakers' oral proficiency, linguistic features of L2 oral production, and their cognitive individual differences in WM and aptitude. The purpose of the dissertation is to investigate whether complexity, accuracy, and fluency (CAF) measures of monologic and dialogic oral production are predictive of longitudinal development in L2 oral proficiency. The dissertation also examines the effects of ESL speakers' individual differences in WM and aptitude on their oral proficiency and how these individual difference variables interact with CAF measures in their effects on L2 oral proficiency.

In second language acquisition (SLA) research, complexity, accuracy, and fluency (CAF) features have been investigated as distinct dimensions of L2 oral and written production (Housen et al., 2012; Norris & Ortega, 2009). Theoretically, the three dimensions of CAF imply major stages in underlying L2 system: development of elaborate and sophisticated L2 knowledge (or higher complexity), restructuring and fine-tuning of L2 knowledge including the non-target-like aspects of interlanguage (higher accuracy), and consolidation and automatization of L2 knowledge with greater performance control (higher fluency) (Housen et al., 2012). Over the last few decades, CAF features of L2 written and oral performances received an increasing amount of attention in SLA literature (Ortega, 2012). According to Ortega (2012), however, less is known about the relationship between CAF features and L2 oral proficiency compared to that between CAF features and L2 written proficiency. The relationships between CAF measures and L2 oral proficiency might vary depending on variables such as task-type (e.g., monologic versus dialogic) and time. Existing research examining the relationships between CAF features and L2 oral proficiency mostly used monologic oral production (e.g., Iwashita et al., 2008; Révész et al., 2016), and some studies used only dialogic oral production (e.g., Tonkyn, 2012). However, few studies examined whether the relationships between CAF measures and L2 oral proficiency vary depending on task-type (monologic versus dialogic). Such an investigation would offer insights into the importance of monologic versus dialogic mode of tasks on linguistic predictors of L2 oral proficiency. Additionally, L2 speakers usually develop their L2 competence over time and there are studies that investigated longitudinal development in CAF measures over time. For example, Tonkyn (2012) examined development in CAF measures over nine weeks and

Vercellotti (2017), over 10 months. However, little is known about how the relationships between CAF measures and L2 oral proficiency vary over time.

In addition to objective CAF measures, variables related to ESL speakers' individual cognitive differences might also be significantly related to their oral skills. Among the cognitive variables, WM and aptitude have been widely studied in relation to L2 oral production features. WM, a system for temporary storage and processing of information during higher order cognitive tasks (Baddeley & Logie, 1999), have been found to be closely related to complex cognitive processes including learning and using a language (Engle, 2002; Juffs & Harrington, 2011; Linck et al., 2014). Additionally, aptitude is traditionally defined as "the ease and speed with which one learns a foreign language" (Li, 2016, p. 804). SLA researchers, for the past two decades, have shown increasing interests in investigating how individual differences in WM and aptitude are related to L2 learning and use (Juffs & Harrington, 2011; Linck et al., 2014; Li, 2016, 2019). Previous research found components of WM (e.g., phonological memory, executive working memory) to be varied but significant predictors of different CAF measures of L2 oral production (e.g., Ahmadian, 2012; Gilabert & Muñoz, 2010). Likewise, Li (2016) found overall aptitude to be a strong predictor of L2 speaking skills. Granena, (2018) also found implicit learning ability (i.e., implicit aptitude) to be a significant predictor of higher fluency in L2 speech. While existing studies mostly examined the relationships between L2 speakers' WM and/or aptitude and CAF features of their oral production, there has not been enough investigations into how WM and aptitude are related to L2 speakers' oral proficiency development over time.

Additionally, individual difference variables (e.g., explicit aptitude or EWM) might interact with different structural features (e.g., CAF measures) that are predictive of L2 oral proficiency (DeKeyser, 2012). For example, a complexity or fluency measure might be significant predictor of L2 oral proficiency only for L2 speakers with higher WM or higher aptitude abilities. Examining such interactions between linguistic structures and individual difference variables (e.g., WM/aptitude) in their combined effects on L2 oral proficiency may suggest why the interacting variables are important in the process of producing efficient speech (DeKeyser, 2012). However, there has been a lack of research examining the interactions between linguistic features of L2 oral production and L2 speakers' individual difference variables (WM/aptitude) in their effects on L2 oral proficiency (DeKeyser, 2012).

Thus, in the literature, there is a lack of clear picture on how the relationships between CAF measures and L2 oral proficiency vary depending on monologic versus dialogic task type, time, or ESL speakers' WM and aptitude abilities. To address these gaps, the dissertation study examines longitudinal development in ESL speakers' oral proficiency in relation to CAF measures of their monologic and dialogic oral production and their individual differences in WM and aptitude. The dissertation has three main research questions. The first question is related to how CAF measures of monologic and dialogic oral tasks are related to L2 oral proficiency, and whether those relationships vary over time. The second research question examines the relationships between WM and aptitude measures and L2 oral proficiency over time. The third research question investigates the interactions between the WM/aptitude measures and the CAF indices in their effects on L2 oral proficiency. As developing L2 speaking skills is important for ESL speakers' academic and social adjustments in higher education contexts (Andrade, 2006, 2009), the results of the dissertation might have important theoretical and pedagogical implications. The results might offer empirically based insights into the linguistic and individual difference variables related to ESL speakers' oral proficiency development in academic contexts.

1.1 Organization of the dissertation

The dissertation is organized in six chapters. Chapter 1 (the present chapter) introduces the main constructs of the dissertation, briefly summarizes the major research gaps, and states the overall purpose of the study. Chapter 2 is literature review, and this chapter is divided into five sections. Each section focuses on a focal construct of the dissertation study. The first section discusses the construct of L2 oral proficiency. The second section discusses monologic and dialogic oral tasks. The third section discusses CAF features of L2 oral production in monologic and dialogic tasks. The fourth section discusses the construct of WM and the fifth section discusses aptitude. The sixth section discusses the motivation for the dissertation and the research questions. Chapter 3 discusses the research methods of the study. Chapter 4 reports the results of the research questions. Chapter 5 discusses the findings, the implications, the limitations of the dissertation study, and some directions for future research. Chapter 6 includes the conclusion.

2 LITERATURE REVIEW

2.1 L2 Oral proficiency

The construct of language proficiency is fundamental in understanding L2 acquisition (Hulstijn, 2012). In literature, there have been many definitions of L2 proficiency, each one tied to a theoretical stance (Leclercq & Edmonds, 2014). Some competing theories of L2 proficiency proposed in SLA literature include cognitive/academic language proficiency (CALP) and basic interpersonal skills (BICS) by Cummins (1980, 1981), communicative competence as grammatical, sociolinguistic, and strategic competence by Canale and Swain (1980), and organizational and pragmatic competence subsuming grammatical, discourse, illocutionary, and sociolinguistic competence by Bachman and Palmer (1985). Hence, L2 oral proficiency as a construct not only means the knowledge of grammar and vocabulary of a language, but it also means the ability to communicate appropriately in the target language (Ortega, 2003). According to Thomas (1994), proficiency corresponds to "a person's overall competence and ability to perform in L2" (p. 330). Hulstijn (2011) provides a more comprehensive definition of L2 proficiency that includes linguistic as well as cognitive competences. According to Hulstijn (2011):

> ". . . language proficiency is the extent to which an individual possesses the linguistic cognition necessary to function in a given communicative situation, in a given modality (listening, speaking, reading, or writing). Linguistic cognition is the combination of the representation of linguistic information (knowledge of form-meaning mappings) and the ease with which linguistic information can be processed (skill). Form-meaning mappings pertain to both the literal and pragmatic meanings of forms (in decontextualized and socially-situated language use, respectively)" (p. 242).

Thus, based on Hulstijn's (2011) definition, the construct of oral proficiency includes not only literal or decontextualized knowledge of language forms but also the pragmatic meanings of those forms, i.e. socially situated use of language.

2.2 Operationalizations of L2 oral proficiency in SLA research

Although L2 proficiency is a fundamental construct in SLA, the way this construct has been measured is not consistent (Thomas, 1994, 2006). Such lack of consistency might be partly due to the context-specific, multidimensional, and multicomponential nature of the oral proficiency construct (Housen & Kuiken, 2009; Thomas, 2006). In SLA-based studies, ESL speakers' oral proficiency has often been operationalized using standardized test scores, for example, ACTFL (American Council on the Teaching of Foreign Languages) (ACTFL, 1999) Oral Proficiency Interviews (OPI) scores (e.g., Halleck, 1995) and TOEFL (Test of English as a Foreign Language) iBT (internet based test) speaking test scores (Iwashita et al., 2008). Oral proficiency has also been operationalized in SLA literature as elicited imitation test scores (Bowden, 2016; Cox & Davies, 2012) and more recently, as functional or communicative adequacy ratings (DeJong et al., 2012a, 2012b; Révész et al., 2016).

2.2.1 The ACTFL OPI

The ACTFL OPI assesses spontaneous, unrehearsed speech and the ability to speak appropriately and effectively in real-life situations (ACTFL, 2020 November 4). The OPI consists of a 20-30 minutes interactive and speaker-centered one-on-one interview between a certified ACTFL tester and a test-taker. ACTFL is a criterion-referenced test, and a test-taker's performance is evaluated with reference to a set of established criteria (e.g., 'Distinguished', 'Superior', 'Advanced', 'Intermediate', 'Novice' etc.). There have been SLA-based studies that used ACTFL OPI for assessing L2 oral proficiency (e.g., Halleck, 1995; Simpson, 2006; Tominaga, 2013). For example, Halleck (1995) investigated the relationships between syntactic complexity indices and L2 oral proficiency levels measured by the ACTFL OPI. The OPI has also been used for implementing and evaluating foreign language programs in the USA (Tominaga, 2013).

2.2.2 TOEFL iBT speaking test

TOEFL iBT is mainly an academic test that measures test-takers ability to "use and understand English at the university level" (ETS TOEFL, 2017b). In the speaking part of TOEFL iBT, test-takers perform two independent speaking tasks followed by four integrated tasks (ETS

TOEFL, 2010)[1]. In the independent tasks, test-takers independently express opinions on familiar topics, whereas in the integrated tasks, test-takers first read and/or listen and then speak. In two of the four integrated tasks, test-takers respond to both oral and written stimuli, whereas in the other two, they respond to only oral stimuli, and the topics are related to both campus situation and academic courses (ETS TOEFL, 2010). TOEFL iBT speaking test is computer administered, and each administration of the test takes 20 minutes (ETS TOEFL, 2010). Iwashita et al. (2008) investigated linguistic features (including various CAF measures) of ESL learners' oral performances underlying the global TOEFL iBT ratings of their oral proficiency. Additionally, Crossley and McNamara (2013) examined whether human judgements of TOEFL iBT speaking scores are predicted by automated linguistic indices, such as those related to delivery (i.e., number of words/ideas), use of language (i.e., grammar, vocabulary), and topic development (i.e., content relevance, coherence). Although TOEFL iBT speaking has high validity and reliability as a standardized proficiency test, administering this test may not be cost-effective.

2.2.3 Elicited imitation test (EIT)

In an EIT, participants listen to a stimulus and then repeat it as exactly as possible (Kim et al., 2016; Tracy-Ventura et al, 2014). EIT has been used as a measure of global L2 oral proficiency (Cox & Davies, 2012; Solon et al., 2019; Tracey-Ventura et al., 2014; Wu & Ortega, 2013). The rationale behind EIT as a measure of oral proficiency is that learners can accurately imitate sentences only if they have comprehended and parsed those through their developing grammars (Tracy-Ventura et al, 2014). In an EIT, repetition of oral sentences measures the

[1] Since August 2019, TOEFL iBT speaking test has been shortened. In the new format, test-takers respond to 4 speaking tasks instead of 6 ("TOEFL Resources", 2020). In the new format, one independent speaking task (task 1: personal preference) and one integrated task (task 5: campus situation) from the old format have been deleted, and the remaining questions are the same as before. During the data collection, authentic TOEFL iBT tests in the new format were not available. Hence, the dissertation study used older versions of the TOEFL iBT speaking test consisting of all the 6 tasks.

ability to comprehend language receptively and productively, to integrate memory of sentences with knowledge of language system from long-term memory, and to employ psychomotor skills necessary for meaningful speech production in real time (Wu & Ortega, 2013).

Ortega et al. (1999) developed EIT in four different languages to examine how syntactic complexity measures were related to L2 oral proficiency cross-linguistically, and they found high reliability, discrimination, and concurrent validity for EIT data (Tracy-Ventura et al., 2014). In addition to its high validity and reliability, EIT does not take much time to administer. It is cost-effective and is available in different languages (Solon et al., 2019). Due to such advantages, EIT has been used for measuring oral proficiency in SLA research (Bowden, 2016; Solon et al., 2019). EIT has also been found to be correlated with various proficiency measures, for example, with OPI in Bowden (2016) and with CAF measures of monologic oral narratives in Wu & Ortega (2013).

2.2.4 Communicative adequacy

Communicative adequacy (also known as “functional adequacy,” De Jong et al., 2012a, 2012b; Kuiken & Vedder, 2018) is defined as "the degree to which a learner's performance is more or less successful in achieving a task's goals efficiently" (Pallotti, 2009, p. 596). Communicative adequacy taps into L2 speakers’ ability to use language appropriately in communicative situations. Kuiken and Vedder (2018) examined communicative adequacy as a component of L2 pragmatics, as the "appropriateness and felicity of the utterances of the speaker/writer within a particular context" (p. 265). In the communicative or functional adequacy rating scale proposed by Kuiken and Vedder (2018), the descriptors are objective, countable, independent from CAF measures, and can be rated by either expert or non-expert raters. For rating communicative adequacy of L2 speech, Kuiken and Vedder (2018) specified four components: content (whether the information units, ideas, or concepts in the speech are adequate and relevant), task requirement (whether the task-requirements have been successfully met with regard to genre and speech act), comprehensibility (how much effort is needed for a listener to understand the purpose and ideas of the speech), and coherence and cohesion (whether the speech is coherent and cohesive).

Kuiken and Vedder (2018) also investigated the efficacy of their communicative adequacy rating scale. In that scale, oral/written performances received separate ratings on a six-point rating scale for each of the four components of the construct: content, task requirement, comprehensibility, and coherence and cohesion. For examining the reliability and validity of this rating scale, Kuiken and Vedder (2018) conducted a study on Dutch L2 and Italian L2 learners who produced two argumentative written texts and performed two oral tasks on the same topics. and four non-expert raters were appointed to rate both the written and spoken data. The results showed significant correlations (ranging from moderate to strong) between the average ratings of the raters on each of the four dimensions of the rating scale. The results also indicate significant and high correlations between the raters' scores on the two tasks (both spoken and written) for all four dimensions of the scale, which indicate that raters judged both texts in similar ways (Kuiken & Vedder, 2018).

Among the studies that employed communicative adequacy for operationalizing L2 oral proficiency, De Jong et al. (2012b) examined how L2 learners’ linguistic knowledge (e.g., knowledge of vocabulary and grammar, formulation and articulation of speech plan) are related to their "success in conveying information through speaking" i.e. communicative adequacy (p. 9). In De Jong et al. (2012b), computer-assisted monologic tasks were used to assess communicative adequacy of the participants' speech. Thus, the participants did not interact with any partner in the tasks although the tasks were fully contextualized with the addressee and the communicative settings specified. The participants had to imagine an audience for the tasks and role-play accordingly. Similarly, Révész et al. (2016) investigated what linguistic factors (such as CAF) are related to communicative adequacy of speaking tasks at various proficiency levels. The participants in Révész et al. (2016) performed five monologic oral tasks fulfilling different functions, such as complaining about a catering service, telling a story based on pictures, giving advice based on an aural commentary, refusing a suggestion, and summarizing information. Similar to De Jong et al. (2012b), Révész et al. (2016) also used computer-delivered speaking tasks. Révész et al. (2016) found that lower pause frequency, a feature of breakdown fluency, and for advanced speakers, the incidence of lower false starts were the strongest predictors of higher communicative adequacy scores.

However, the above-mentioned studies examined communicative adequacy in only monologic oral tasks. Although those tasks served a functional purpose, they did not involve real interactions between participants. Communicative adequacy is related to the idea of interactional competence or "what a person does together with others" (Pallotti, 2009; Young, 2011, p. 430). To produce a functionally effective speech, "participants recognize and respond to the expectations of what to say and how to say it, contingent on what other participants do and what the context is" (Révész et al., 2016, p. 830). Therefore, speakers’ success in fulfilling the communicative requirements of an oral task might be dependent on the monologic versus dialogic nature of the task, and hence, communicative adequacy needs to be measured for both monologic and dialogic oral tasks. However, this issue has not received much attention in the research literature.

2.3 Monologic and dialogic oral tasks

Depending on discourse mode, oral tasks can be either monologic or dialogic (R. Ellis, 2003). The focus of this dissertation is on comparing CAF-based measures of monologic and dialogic tasks as predictors of L2 oral proficiency. In monologic tasks, no interlocutors are involved while in dialogic tasks, speakers interact with at least one interlocutor. . Table 2.1 summarizes how the task features introduced in R. Ellis (2003, 2012) are related to both monologic and dialogic tasks.

Table 2.1 Features of Monologic and Dialogic Speaking Tasks (R. Ellis, 2003, 2012)

Feature	Relevant to Monologic Task	Relevant to Dialogic Task
➢ Unfocused (eliciting target language in general) versus focused (eliciting specific target forms)	Yes	Yes
➢ Input-providing (engaging learners in listening/reading) versus output-prompting (engaging learners in speaking/writing)	Yes	Yes
➢ Closed (with a single possible outcome) versus open (with multiple possible outcomes)	Yes	Yes
➢ Structured versus unstructured	Yes	Yes
➢ Cognitive processes	e.g., Tasks involving reasoning, expression of personal opinions, narration	e.g., Tasks involving discussion on argumentative topics, information exchange
➢ Presence of a partner and production of interactive speech	N\A	Yes
➢ Goal orientation: Either the task requires the participants to agree on a single outcome (convergent) or the task allows them to disagree (divergent)	N\A	Yes

In a monologic task, a participant individually delivers a narrative, and there is no interaction involved with any partner or interlocutor (Skehan, 2001). In contrast, dialogic or interactive tasks "require interaction, and a discourse style that leads participants to alternate in who holds the floor" (Skehan, 2001, p. 173). In an interactive discourse, speakers need to connect their utterances with those from their interlocutors by using various turn-opener tokens that helps flow their conversations better (McCarthy, 2010). However, such requirements are not present in monologic oral performances. Thus, from a pragmatic perspective, dialogic speech might be more complex than monologic speech (Michel, 2011). In L2 research, learners' performances in monologic and dialogic oral tasks have been measured using CAF constructs.

2.4 CAF Features of oral production

Oral production features (i.e., CAF measures) are distinct from oral proficiency (Ortega, 2012). According to Housen et al. (2012), CAF are "multilayered, multifaceted, and multidimensional constructs" (p. 5). CAF measures have been used in task-based research not only as linguistic features of L2 oral production (Ortega, 2012) but also as indicators of L2 proficiency underlying that production (Granena, 2018; Housen & Kuiken, 2009). Skehan (1989) first proposed an L2 model that included CAF as the three principal dimensions of L2 proficiency. In traditional definitions, complexity is "the extent to which the language produced in performing a task is elaborate and varied" (R. Ellis, 2003, p. 340). Accuracy is the ability to produce speech free of errors (Housen & Kuiken, 2009), and fluency refers to "the extent to which the language produced in performing a task manifests pausing, hesitation, or reformulation" (R. Ellis, 2003, p. 342). As argued by Housen and Kuiken (2009), complexity and accuracy are related to L2 knowledge representation while fluency is related to "learners' control over their linguistic L2 knowledge" (p. 462). Considering the multidimensional nature of the CAF constructs, Norris and Ortega (2009) emphasize examining CAF as a dynamic and interconnected group of continuously changing systems. The interconnections among the CAF measures might take various forms. For example, increase in fluency may occur at the expense of increase in accuracy and complexity (R. Ellis, 2008). CAF measures might also vary depending on the variations in oral task-types (e.g., monologic and dialogic) (De Jong et al., 2012a; R. Ellis, 2012; Robinson, 2001a; Skehan, 2001; Tavakoli, 2016). The following sub-sections discuss the theoretical definitions of the CAF constructs as well as the empirical studies examining these constructs.

2.4.1 Syntactic complexity

In literature, there have been several perspectives on the definition of complexity, for example, linguistic complexity, cognitive complexity, and developmental complexity (Michel, 2017). Linguistic complexity refers to "intrinsic formal or semantic-functional properties of L2 elements (e.g., forms, meanings, and form-meaning mappings)" (Housen et al., 2012, p. 4). Additionally, cognitive complexity refers to the "subjective difficulty of a language feature, that is, how a learner perceives the difficulty of an item as it is processed and learned" (Michel, 2017, p. 52). Moreover, developmental complexity refers to "the order in which linguistic structures emerge and are mastered in second (and possibly, first) language acquisition" (Pallotti, 2015, p. 118). The present study focuses on linguistic complexity.

In theoretical level, linguistic or grammatical complexity is defined by Bulté and Housen (2012) as structural complexity that refers to the depth or embeddedness of L2 forms. Additionally, as a behavioral construct, Bulté and Housen (2012) defined complexity as grammatical diversity (including complexity at sentence, clausal, and phrasal level) and as grammatical sophistication (including morphological complexity, both inflectional and derivational). The dissertation study focuses on complexity as grammatical diversity (Bulté & Housen, 2012) because the study measures complexity at sentence, clausal, and phrasal levels.

The importance of different complexity measures might be dependent on the proficiency levels of speakers and the type of content that they produce in different tasks in different modalities (oral/written) (Ortega, 2012). Therefore, a single complexity measure might not be a reliable indicator of proficiency at all levels: beginner, intermediate, and advanced (Ortega, 2012). At the beginner level of L2 development, syntactic complexity is characterized mainly by clausal co-ordination, which has been rarely investigated in CAF-based studies in SLA (Norris & Ortega, 2009). At the intermediate level, subordination-based measures can indicate increase in syntactic complexity (Norris & Ortega, 2009; Ortega, 2012). However, as learners move past the intermediate level, they tend to asymptote in terms of clausal subordination and move more toward phrasal complexity (Norris & Ortega, 2009). Hence, mean length of clause, that measures complexification at the phrasal level and is not influenced by the amount of subordination, is proposed as "a good global index of complexity" for languages typically produced by advanced and matured learners in formal academic contexts (Ortega, 2012, p. 145). It has been argued that in addition to overall sentence complexity and subordination measures, SLA studies investigating L2 complexity should also include measures of coordination and phrasal complexity (Bulté & Housen, 2012; Norris & Ortega, 2009). However, in CAF studies, sub-sentential clausal or phrasal complexity measures (e.g.,, mean length of clauses, mean length of noun/verb phrases) and specific features of L2 knowledge system (e.g., frequencies of different grammatical forms) received less attention compared to general or global complexity indices (e.g., mean length of analysis of speech [AS][2] unit, mean length of T[3]-unit) (Bulté & Housen,

[2] AS unit refers to an independent clause or sub-clausal unit together with any subordinate clause associated with either (Foster et al., 2000).
[3] T-unit refers to one main clause and any subordinate clause attached to that main clause (Hunt, 1966).

2012; Robinson & N. Ellis, 2008). Some recent studies examined clausal and phrasal complexity measures in L2 writing (e.g., Kyle & Crossley, 2018; Staples et al., 2016).

2.4.2 Empirical studies on syntactic complexity in L2 oral production

Whereas CAF measures have been widely investigated in task-based SLA studies, only a few studies examined syntactic complexity in both monologic and dialogic oral tasks. Michel et al. (2007) and Michel (2011) examined the effects of the variations between monologic and dialogic task types on the oral performances of L2 learners of Dutch and found that the participants produced less complex language in dialogic speech compared to that in monologic speech. Similarly, Ferrari (2012) investigated longitudinal development in CAF features in monologic and dialogic speech of four Italian as L2 learners and found that the participants produced longer clauses and more complex AS-units in monologic tasks than in dialogic tasks.

Additionally, the studies that examined the relationships between complexity measures and L2 oral proficiency used either monologic or dialogic task-types. For example, Révész et al. (2016) investigated the relationship between CAF measures of ESL learners' monologic oral tasks and their communicative adequacy scores and found that learners with higher communicative adequacy produced more complex subordinate and conjoined clauses. Furthermore, Iwashita et al. (2008) investigated features of ESL learners' oral performances underlying the global ratings of their TOEFL iBT speaking scores. Iwashita et al. (2008) found that the participants with higher oral proficiency produced more complex verb-phrases and longer utterances. Moreover, Tonkyn (2012) examined the relationships between CAF measures and subjective ratings of ESL learners' speech in dialogic interview data. Tonkyn (2012) found that the participants with higher oral proficiency used longer AS-units, more subordinate clauses, and more primary auxiliaries. Similarly, Halleck (1995) investigated the relationships between the ACTFL oral proficiency levels and syntactic complexity measures in the dialogic OPI data of 107 English as foreign language learners. It was found that the participants at the “superior” level produced more syntactically complex language than those at the “advanced” and “intermediate” levels (Halleck, 1995). Thus, previous studies found that L2 speakers of higher oral proficiency produced syntactically more complex language.

Several studies in literature also examined longitudinal development in CAF measures of oral production. Tonkyn (2012) examined changes in CAF measures in the oral interview data of upper intermediate level instructed learners of English over nine weeks. Tonkyn (2012) found that several syntactic complexity measures, for example, number of subordinate clauses, modal and catenative verbs, and the use of adverb-based adverbials showed significant progress over time. Similarly, Vercellotti (2017) examined development in CAF measures in English language learners' monologic oral performances over 10 months. Vercellotti (2017) found that for participants with higher initial proficiency, their mean length of AS-unit significantly increased over time. Additionally, Ferrari (2012) examined development in CAF features in the monologic and dialogic oral data of four Italian as L2 learners over three years and found that over time, the participants’ scores significantly increased for clause lengths, but not for subordination. The findings of Ferrari (2012) support the argument (Ortega, 2003, 2012) that as L2 learners advance in proficiency over time, complexification measures at the phrasal level become more important than subordination.

Overall, previous studies found development in both general (e.g., mean length of AS-unit in Vercellotti, 2017) and specific (e.g., number of modal and catenative verbs in Tonkyn, 2012) syntactic complexity measures over time for ESL learners of varied proficiency levels. However, as these studies did not examine the relationships between the complexity indices and oral proficiency, it is not clear whether the syntactic complexity measures were also predictive of development in L2 oral proficiency over time. Additionally, although previous studies (e.g., Iwashita et al., 2008; Tonkyn, 2012) showed that ESL speakers of higher oral proficiency produced more complex language, it remains under-explored whether such relationships between syntactic complexity and L2 oral proficiency vary depending on the monologic versus dialogic nature of the speaking tasks.

2.4.3 Lexical sophistication

Because of the crucial role played by lexis alongside syntax in speech production (Levelt, 1989, 1999), lexical complexity has been examined alongside syntactic complexity in L2 acquisition research (Skehan, 2009). At the theoretical level, Bulté and Housen (2012) defined lexical complexity as systemic lexical complexity (i.e., elaboration, range, size, and breadth of L2 lexical items) and structural lexical complexity (i.e., depth of L2 lexical items). Additionally, lexical density (e.g., lexical words/function words) and diversity (e.g., type/token ratios, number of word type) measures tap into systemic lexical complexity while lexical sophistication measures (e.g., frequency-based type/token ratios) tap into structural lexical complexity (Bulté & Housen, 2012). The present study focuses on lexical sophistication that refers to the use of "advanced vocabulary" in terms of both the depth and breadth of lexical production (Bardel et al., 2012, p. 270). In literature, there has been considerable focus on lexical diversity measures (such as type-token ratio that tap into the breadth of lexical knowledge) for measuring lexical complexity in L2 monologic and dialogic tasks (e.g. Michel et al., 2007; Michel, 2011). Studies that measured lexical sophistication (Gass et el., 1999; Iwashita et al., 2008) mostly focused on frequency-based indices such as the frequency of word types and tokens per minute in monologic speech in Iwashita et al. (2008).

However, such lexical diversity and frequency-based indices draw on surface level lexical features and may not capture the depth of L2 speakers' lexical knowledge such as their knowledge of semantic relations of L2 words (Salsbury et al., 2011). In recent years, advances in computational linguistics and the development of natural language processing (NLP) tools for automatic analysis of lexical diversity and sophistication using large corpora (Graesser et al., 2004; Kyle & Crossley, 2014) have made it possible for researchers to investigate the conceptual development of word knowledge among L2 learners that go beyond the traditional type/token ratio or frequency-based indices (Crossley et al., 2009; Salsbury et al., 2011). The current study uses a computational tool, TAALES (Tool for the Automatic Analysis of Lexical Sophistication) by Kyle and Crossley (2014) for measuring lexical sophistication of L2 oral performances. Among the lexical sophistication indices measured by TAALES, the dissertation focuses on the psycholinguistic word information indices based on the Medical Research Council database (Coltheart, 1981) and the spoken frequency measure based on the Corpus of Contemporary American English (COCA) (Davis, 2008).

2.4.4 Empirical studies on lexical complexity and sophistication in L2 oral production

Michel et al. (2007) and Michel (2011) compared lexical complexity between monologic and dialogic tasks and found that L2 learners of Dutch had higher lexical complexity (higher percentage of lexical words) in dialogic tasks compared to that in the monologic tasks. Additionally, among the studies that investigated development of CAF measures over time, Vercellotti (2017) found that ESL learners with higher initial proficiency had higher lexical diversity (type/token ratio) over time, although the pattern of growth was non-linear (with a dip followed by a steeper increase). Likewise, in Tonkyn (2012), ESL learners’ use of rare word-types significantly increased over time.

Moreover, among the studies that examined the relationship between CAF measures and L2 oral proficiency, Iwashita et al. (2008) found that ESL speakers of higher oral proficiency produced a wider range of word types in monologic speech. Likewise, in Révész et al. (2016), ESL learners who received higher communicative adequacy ratings produced lexically more diverse words in monologic tasks. Tonkyn (2012) also found that ESL learners with higher oral proficiency used higher number of word types than those with lower proficiency in dialogic interviews. Thus, previous findings showed that ESL speakers with higher oral proficiency generally used more diverse vocabulary.

In literature, few studies used NLP tools for examining lexical sophistication in L2 oral performances. Salsbury et al. (2011) analyzed lexical development in the spoken data of six adult ESL learners in a one-year longitudinal study. Salsbury et al. (2011) used word information scores from the Medical Research Council (MRC) psycholinguistic database (Coltheart, 1981) to examine L2 learners' depth of word knowledge measured by psycholinguistic values for concreteness (the extent to which a word refers to an object, material, or person), imageability (whether a word has a strong or weak image related to it), meaningfulness (how related a word is to other words), and familiarity (how familiar to adults a word is). The results showed that the L2 learners’ vocabulary became less context-dependent, more abstract, and more tightly associated over time. Additionally, Kyle and Crossley (2015) used TAALES to examine a wide range of lexical sophistication indices related to frequency, range, academic language, and psycholinguistic word information in L2 spoken data and found that ESL speakers of higher oral proficiency used less familiar and more academic words and more frequent content words.

Overall, majority of studies examining lexical complexity in L2 oral production focused on lexical diversity measures (e.g., type-token ratio) (e.g., Iwashita et al., 2008; Tonkyn, 2012; Vercellotti, 2017). However, such indices may not tap into the depth of ESL speakers' lexical knowledge (Salsbury et al., 2011). Additionally, the studies examining the relationships between lexical complexity and oral proficiency focused on either monologic or dialogic data. Hence, there is a lack of investigations into how the variations between monologic versus dialogic task types affect the relationships between lexical sophistication and L2 oral proficiency and whether this relationship changes over time.

2.4.5 Accuracy

Of the CAF triad, accuracy is the most straightforward and consistent construct (Housen et al., 2012; Pallotti, 2009). According to Housen et al. (2012), accuracy refers to "the extent to which an L2 learner's performance (and the L2 system that underlies this performance) deviates from a norm," and such deviations are traditionally labelled as errors (p. 4). Increasing accuracy in L2 production is one feature of L2 acquisition (Foster & Wigglesworth, 2016). Accuracy can be measured locally or globally. Local measures of accuracy count the accurate use of specific L2 grammatical features (for example, verb and noun morphology), whereas global measures calculate the overall accuracy in L2 performance (Foster & Wigglesworth, 2016). In the research domain of L2 oral production, different methods, both local and global, have been used to measure accuracy. For example, Michel (2011) examined lexical errors, morpho-syntactic errors, and determiner errors in monologic and dialogic oral tasks. Additionally, global measures such as percentage of error-free syntactic units (for example, error free AS-unit in Tonkyn, 2012 and Ferrari, 2012, percentage of error-free T-unit in Iwashita et al., 2008 and error-free clauses in Vercellotti, 2017) were commonly used accuracy measures across monologic and dialogic task types in previous studies. L2 studies usually measured accuracy for errors in both lexis and syntax (e.g., Révész et al., 2016).

2.4.6 Empirical studies on accuracy in L2 oral production

In Michel et al. (2007) and Michel (2011), L2 learners of Dutch were significantly more accurate in dialogic tasks than in monologic tasks in the following measures: total number of syntactic errors, lexical errors, and omissions per AS unit. Several studies also examined the relationships between accuracy of oral performances and L2 oral proficiency. In Révész et al. (2016), number of errors per 100 words in monologic speech was a significant predictor of ESL learners' communicative adequacy. Likewise, in Tonkyn (2012), ESL learners with higher oral proficiency produced more accurate verb phrases in dialogic interviews. Similarly, Iwashita et al. (2008) found that grammatical accuracy measures in TOEFL task performances were significant predictors of TOEFL iBT oral proficiency scores.

Additionally, several studies that examined longitudinal development in CAF measures found significant development in L2 accuracy scores over time. For example, in Vercellotti (2017), ESL learners with higher initial proficiency had a linear growth in the percentage of error-free clauses over 10-months. Likewise, in Tonkyn (2012), ESL learners had significant gains in producing longer accurate syntactic units and accurate noun phrase and verb phrases over time in dialogic tasks. Similarly, in Ferrari (2012), Italian as L2 learners' accuracy rate (percentage of error-free AS-units) in monologic and dialogic speech increased over three years although the pattern of the development was non-linear with an initial decrease in accuracy followed by an increase.

Overall, previous findings showed that L2 speakers of higher oral proficiency produced more accurate speech and that the rate of accuracy in L2 speech also increased over time.

However, few studies examined accuracy in both monologic and dialogic speech as predictors of L2 oral proficiency. Additionally, there is a lack of investigations into accuracy as a predictor of oral proficiency development over time. Hence, it remains understudied whether the relationships between accuracy and L2 oral proficiency varies depending on task-type (e.g., monologic versus dialogic) and time.

2.4.7 Oral fluency

In its broad definition, fluency refers to overall oral proficiency (Housen et al., 2012; Huensch & Tracy-Ventura, 2017). However, in its narrow definition, fluency refers to “the temporal aspects of oral production that influence the degree of fluidity in speech (e.g. pauses, hesitations, speech rate)” (Derwing et al., 2009, p. 534). The present study adopts this narrow definition of fluency. Segalowitz (2010) defined three different types of fluency: cognitive fluency, utterance fluency, and perceived fluency.

Cognitive fluency refers to a speaker's ability to "efficiently mobilize and integrate the underlying cognitive processes responsible for producing utterances” (Segalowitz, 2010, p. 48). Additionally, utterance fluency refers to the features of an utterance, i.e., the temporal, pausing, hesitation, and repair characteristics (Segalowitz, 2010). In contrast to cognitive fluency that is concerned with a speaker's internal cognitive abilities, utterance fluency refers to the fluency characteristics of a sample of speech (Segalowitz, 2010). Furthermore, perceived fluency refers to the judgement that listeners make about speakers based on the impressions drawn from their samples of speech (Segalowitz, 2010). Similar to the majority of studies on L2 fluency, the present study focuses on utterance fluency (Huensch & Tracy-Ventura, 2017).

As fluency is a multidimensional and multifaceted construct (Housen et al., 2012; Tavakoli & Skehan, 2005), Tavakoli and Skehan (2005) further divided utterance fluency into three sub-dimensions: breakdown fluency, speed fluency, and repair fluency. Breakdown fluency is concerned with silence (Tavakoli & Skehan, 2005). It refers to the duration and number of pauses and lengths of runs. Some common indices used to measure breakdown fluency include length and number of unfilled and filled pauses, total duration of silence (Tavakoli & Skehan, 2005), and length of run (De Jong et al., 2012a). Another measure of breakdown fluency, phonation-time ratio, summarizes the measures related to pausing because this measure indicates the percentage of time filled only with speech (total length of speech or phonation divided by the total utterance time) (De Jong et al., 2012a). Additionally, speed fluency refers to the "speed with which language is produced" (Tavakoli & Skehan, 2005, p. 254). Speed fluency is usually measured by counting the number of words or syllables produced per time unit, for example, articulation rate and speech rate (De Jong et al., 2012a). Furthermore, repair fluency refers to "reformulation, replacement, false starts, and repetition of words or phrases" (Tavakoli & Skehan, 2005, p. 255). Repair fluency can be measured by counting the number of hesitations and false starts (De Jong et al., 2012a).

2.4.8 Empirical studies on oral fluency in L2 oral production

Several studies in literature compared fluency measures between L2 monologic and dialogic tasks. In Michel et al. (2007) and Michel (2011), L2 learners of Dutch had higher fluency in dialogic tasks than in monologic tasks. In dialogic tasks, learners were significantly more fluent in unpruned speech (including reformulations, repetitions, and replacements) as well as in pruned speech (excluding reformulations, repetitions, and replacements), and they also produced fewer filled pauses (e.g. uhm, mmm) in dialogic speech than in monologic speech

(Michel et al., 2007; Michel, 2011). Likewise, Tavakoli (2016) compared L2 fluency measures between monologic and dialogic speech. In Tavakoli (2016), similar to Michel et al. (2007) and Michel (2011), ESL speakers had higher fluency in dialogic speech than in monologic speech because the participants significantly produced longer fluent runs, shorter pauses, higher phonation time ratio, and faster articulation rates in dialogues than in monologues. Likewise, in Ferrari (2012), L2 learners of Italian had less pauses and hesitations in the dialogic tasks than in the monologic tasks. Overall, previous studies found higher fluency in L2 dialogic speech than in monologic speech.

Among the studies that examined the relationships between fluency measures and L2 oral proficiency, measures of speed and breakdown fluency have often been found to be stronger predictors of L2 oral proficiency than those of repair fluency (Huensch & Tracy-Ventura, 2017). Révész et al. (2016) found that filled pause frequency, a measure of breakdown fluency, was the strongest predictor of ESL speakers' communicative adequacy in monologic speech. ESL learners with higher communicative adequacy produced fewer filled pauses (Révész et al., 2016). Similarly, in Iwashita et al. (2008), fluency-based measures (pause-time and speech-rate) were significantly related to TOEFL iBT oral proficiency ratings. ESL speakers with higher oral proficiency spoke faster with less pausing (Iwashita et al., 2008). In similar vein, Tonkyn (2012) found that ESL speakers with higher oral proficiency took fewer pauses and had higher speech rate (syllables per minute), less false starts and repetitions, and longer fluent runs than low proficient learners. Overall, previous studies found that ESL learners with higher oral proficiency had higher speed and breakdown fluency.

Moreover, there have been longitudinal studies that found development in L2 fluency measures over time. In Vercellotti (2017), ESL learners with higher initial proficiency improved their fluency scores (shorter lengths of pauses) over 10 months. Similarly, in Ferrari (2012), L2 learners of Italian had linear development in fluency scores over three years because both pauses and hesitations in their speech decreased over time. Likewise, in Tonkyn (2012), the length of fluent runs and the length of turns in ESL learners' speech showed significant improvement over nine weeks.

Therefore, previous studies found various L2 fluency measures to develop over time. However, these studies did not examine whether L2 fluency measures are also predictive of development in L2 oral proficiency over time. Such an investigation would offer insights into the role played by fluency indices in longitudinal development in L2 oral proficiency. Additionally, previous studies examined the relationships between fluency measures and L2 oral proficiency either in monologic (e.g., Iwashita et al., 2008; Révész et al., 2016) or in dialogic (Tonkyn, 2012) tasks. Hence, there is no clear picture of whether or how variations in speaking task-type (e.g., monologic versus dialogic) affect the relationships between fluency indices and L2 oral proficiency.

To highlight how CAF constructs have been measured in SLA studies focused on L2 oral production, Table 2.2 shows the operationalizations of CAF constructs in L2 studies that employed monologic and/or dialogic oral tasks. In Table 2.2, the measures of syntactic complexity, lexical complexity, accuracy, and fluency are presented separately under labelled rows. For the studies that examined the relationships between CAF measures and L2 oral proficiency, Table 2.2 also mentions the measures that were significantly related to oral proficiency.

Table 2.2 CAF Measures in L2 Studies Examining Monologic and/or Dialogic Oral Tasks

Measures of Syntactic Complexity			
Study	Monologic Tasks	Dialogic Tasks	The relationship between syntactic complexity and oral proficiency
Halleck (1995)	N/A	--Mean T-unit length, --Mean error-free T-unit length, --Percent of error-free T-units	Longer T-units related to high proficient speech
Michel et al. (2007)	--Total number of clauses per AS-unit, --Ratio of subordinate clauses per total number of clauses	--Total number of clauses per AS-unit, --Ratio of subordinate clauses per total number of clauses	N\A
Iwashita et al. (2008)	--Number of clauses per T-unit, --Ratio of dependent clauses to the total clauses, --Number of verb phrases per T-unit, --Mean length of utterance	N/A	Higher verb-phrase complexity and longer length of utterance predictive of high proficient speech
Tonkyn (2012)	N/A	--Total number of words, --Subordinate clauses, --Primary and modal auxiliaries, --Catenative verbs, --Adverbial adverbs, --Adverbial prepositional phrases	Longer AS-units, more subordinate clauses, and more primary auxiliaries related to high proficient speech
Ferrari (2012)	--Average number of subordinate clauses per AS-unit,	--Average number of subordinate clauses per AS-unit,	N\A

	--Average number of words per clause	--Average number of words per clause	
Révész et al. (2016)	--Subordination measure, --Phrasal complexity, --Overall complexity (ratio of words to AS-units), ---Frequency and Guiraud's index for tense-aspect forms, modal verbs, and types of clauses	N/A	Higher subordination and frequency of conjoined clauses predicted higher communicative adequacy
Vercellotti (2017)	--Mean length of AS-unit	N/A	N\A
Measures of Lexical Complexity			
Study	Monologic Tasks	Dialogic Tasks	The relationship between lexical complexity and oral proficiency
Michel et al. (2007)	--Guiraud’s Index, --Percentage of lexical words	--Guiraud’s Index, --Percentage of lexical words	N/A
Iwashita et al. (2008)	--Proportions of low and high frequency vocabulary (both type and token)	N/A	Wider range of word types predictive of higher oral proficiency
Michel (2011)	--Guiraud’s Index	--Guiraud’s Index	N/A
Tonkyn (2012)	N/A	--Frequency of word types and word families, --Use of less frequent word tokens, word types, and word families	Higher number of word types predictive of higher oral proficiency

Vercellotti (2017)	--Lexical variety calculated as "D"	N/A	N/A

Measures of Accuracy

Study	Monologic Tasks	Dialogic Tasks	The relationship between accuracy and oral proficiency
Michel et al (2007)	--Total number of errors --Lexical errors --Omissions per AS-unit, --Percentage of self-repairs	--Total number of errors --Lexical errors --Omissions per AS-unit, --Percentage of self-repairs	N/A
Iwashita et al. (2008)	--Error free T-units, --Errors in verb tense, third person singular, plural markers, prepositions, and article use	N/A	Grammatical accuracy predicted higher oral proficiency
Michel (2011)	--Lexical errors, --Morpho-syntactic errors --Determiner errors per AS-unit	--Lexical errors, --Morpho-syntactic errors, --Determiner errors per AS-unit	N/A
Tonkyn (2012)	N/A	--Error-free AS-units/Total AS-units, --Words/error-free AS-unit, --Words/verb phrase error, --Words/noun phrase error, --Words/syntactic error, --Words/lexical error	Use of accurate verb phrases predictive of higher oral proficiency
Ferrari (2012)	--Percentage of error-free AS-units	--Percentage of error-free AS-units	N/A

Révész et al. (2016)	--Proportion of errors per 100 words, --Correct use of subject-verb agreement, tense-aspect forms, modal verbs, connectors	N/A	Number of errors per 100 words predictive of communicative adequacy
Vercellotti (2017)	--Percentage of error-free clauses	N/A	N/A
Measures of Fluency			
Study	Monologic Tasks	Dialogic Tasks	The relationship between fluency and oral proficiency
Michel et al. (2007)	--Ratio of syllables per minutes in unpruned and pruned speech, --Number of filled pauses per 100 words	--Ratio of syllables per minutes in unpruned and pruned speech, --Number of filled pauses per 100 words	N/A
Iwashita et al. (2008)	--Filled and unfilled pauses, --Repair by 60 seconds of speech, --Total pausing time, --Speech rate (total syllable/total utterance time), --Mean length of run	N/A	Higher speech rate and fewer pauses predictive of higher oral proficiency
Michel (2011)	--Speech rate (syllables per second) in unpruned and pruned speech, --Repairs per AS-unit, --Filled pauses per AS-unit	--Speech rate (syllables per second) in unpruned and pruned speech, --Repairs per AS-unit, --Filled pauses per AS-unit	N/A
Ferrari (2012)	--Average number of silent pauses per AS-unit, --Average number of hesitation phenomena	--Average number of silent pauses per AS-unit, --Average number of hesitation phenomena	N/A
Tonkyn (2012)	N/A	--Rate of speaking (syllables/minute),	Fewer pauses, higher speech

		--Mean length of fluent runs, --Phonation time/total speaking time, --Proportion of total pause time at text-unit boundaries, --Mean length of turns, --Number of words excluding false starts, --Repetitions/total words, pause clusters	rate, less false starts and repetitions, and longer fluent runs predictive of higher oral proficiency
Révész et al. (2016)	--Silent and filled pauses per 100 words, --Mean duration of syllables, --False starts, self-repairs, and repetitions per 100 words	N/A	Fewer filled pauses predictive of higher communicative adequacy
Tavakoli (2016)	--Articulation rate, --Speech rate, --Mean length of pauses (per 60 seconds), --Mean number of pauses (per 60 seconds), --Mean number of repetitions, hesitations, and false starts, --Mean number of filled pauses, --Mean length of run, --Phonation/time ratio	--Articulation rate, --Speech rate, --Mean length of pauses (per 60 seconds), --Mean number of pauses (per 60 seconds), --Mean number of repetitions, hesitations, and false starts, --Mean number of filled pauses, --Mean length of run, --Phonation/time ratio, --Number of turns, --Number of interruptions	N/A
Vercellotti (2017)	--Mean length of pause	N/A	N/A

As can be seen in Table 2.2, most of the studies used multiple measures for operationalizing each of the CAF constructs, which reflects the multidimensionality of these constructs (Housen et al., 2012). Table 2.2 also shows that the studies that examined the relationships between CAF measures and L2 oral proficiency included either monologic or dialogic tasks, not both. Additionally, the studies that examined the variations in CAF measures between monologic and dialogic tasks found significant differences in the measures between the two task types (except Michel et al., 2007 who did not find any difference for the lexical complexity indices). These findings confirm the theoretical arguments that linguistic features of oral production (e.g., CAF measures) might vary depending on task-type (e.g., monologic versus dialogic) (Robinson, 2001a, 2001b; Skehan, 2001).

Overall, based on the discussions above, three strands of research have been identified on CAF measures in L2 oral production: studies that examined the variations in CAF measures between monologic and dialogic tasks, studies that examined the relationships between CAF measures and L2 oral proficiency, and studies that examined development in CAF measures over time. Table 2.3 summarizes overall findings of these studies on CAF measures. In Table 2.3, each of the above-mentioned strands of research is presented separately under a labelled row. Table 2.3 also mentions the type of monologic and/or dialogic task used in each study.

Table 2.3 Summary of the Empirical Studies Examining CAF Measures in L2 Oral Production

Studies examining variations in CAF measures between monologic and dialogic oral tasks		
Study	Type of Oral Task Used	General Findings
Michel et al. (2007)	Monologic (leaving a phone message to a friend giving advice about which MP3 player or cell phone to buy) Dialogic (doing a phone conversation on the same topic)	--Syntactic complexity higher in monologic tasks --Higher lexical diversity in dialogic tasks -- Higher accuracy in dialogic tasks -- Higher fluency in dialogic tasks

Michel (2011)	Monologic (leaving a phone message to a friend about choosing the best dating or study couple) Dialogic (doing a phone conversation on the same topic)	--Syntactic complexity higher in monologic tasks -- Higher lexical diversity in dialogic tasks --Higher accuracy in dialogic tasks -- Higher fluency in dialogic tasks
Ferrari (2012)	Monologic (film retelling, story retelling) Dialogic (interview and excerpts from initial parts of a telephone conversation)	-- Syntactic complexity higher in monologic speech -- Higher fluency in dialogic speech
Tavakoli (2016) (only focused on fluency)	Monologic (retelling a recent shopping experience) Dialogic (a discussion task arguing for or against a topic, e.g., watching a movie at home or in the cinema)	-- Higher fluency in dialogic speech
Studies examining the relationships between CAF measures and L2 oral proficiency		
Study	Type of Oral Task Used	General Findings
Iwashita et al. (2008)	Monologic (responses to TOEFL iBT speaking test)	-- Complex verb-phrases and longer utterances predictive of higher proficiency -- High proficiency learners used wider range of word types --Grammatical accuracy higher for high proficiency learners -- Less pausing and higher speech-rate predictive of higher oral proficiency
Tonkyn (2012)	Dialogic (interviews with the researcher on academic disciplines and English learning experience)	-- Higher syntactic complexity (e.g., longer AS-unit) predictive of higher oral proficiency -- Higher number of word types used by higher proficiency participants --More accurate verb phrases produced by higher proficiency learners --Higher speech rate and fewer pauses predictive of high proficiency speakers
Révész et al. (2016)	Monologic (tasks with specific communicative functions to fulfil, e.g., summarizing, giving advice, refusing a suggestion etc.)	--Higher subordination and conjoined clauses predictive of higher communicative adequacy -- Higher lexical diversity predictive of higher communicative adequacy

		-- Higher accuracy predictive of higher communicative adequacy -- Speech with lower filled pauses predictive of higher communicative adequacy
Studies examining development in CAF measures over time		
Study	Type of Oral Task Used	General Findings
Tonkyn (2012)	Dialogic (interviews with the researcher on academic discipline and English learning experience)	--Complexity measures (e.g., use of subordination, modal verbs) developed over time -- Use of rare word-types increased over time --Development in accuracy over time -- Length of fluent runs and length of turns increased over time
Ferrari (2012)	Monologic (film retelling, story retelling) Dialogic (interview and excerpts from initial parts of a telephone conversation)	--Clausal complexity developed over time -- Accuracy rate increased over three years -- Pauses and hesitations decreased over time
Vercellotti (2017)	Monologic (monologues on the topics in the Intensive English Program's curriculum)	--Mean length of AS-unit developed over time -- Lexical diversity scores higher over time -- Growth in the accuracy scores over time -- Higher fluency (lower mean pause-lengths) over time

As can be seen in Table 2.3, among the few studies that compared CAF measures between monologic and dialogic tasks, monologic speech had significantly higher syntactic complexity, and dialogic speech had significantly higher fluency. Hence, monologic versus dialogic tasks had varied impacts on the CAF measures of L2 speakers' oral performances (Robinson, 2001; Tavakoli, 2016). However, these empirical findings do not support the theoretical arguments proposed by Skehan (2001) regarding the effects of monologic versus dialogic tasks on CAF measures of L2 oral production. Skehan (2001) argued that in a dialogic task, a participant gets time to focus on accuracy while their partner is speaking, and they can also re-use their partner's language to recycle correct language. Moreover, in such a task, reinterpretation of the task together with the partner as well as explanation of the partner's language may lead to higher complexity, but the need to involve in online planning and the "uncertainty of turn-taking" might lead to reduced fluency (Skehan, 2001, p. 176). Thus, Skehan (2001) argued that compared to monologic tasks, dialogic tasks might have greater accuracy and complexity but lower fluency.

Additionally, as shown in Table 2.3, the existing studies that examined the variations in CAF measures between monologic and dialogic oral tasks (e.g., Ferrari, 2012; Tavakoli, 2016) used different topics for these two task types. Hence, it is not clear to what extent the significant differences in CAF measures between these task types could be attributed to the topic variance. Furthermore, in Ferrari (2012), the monologic task included retellings of films and stories that might not elicit spontaneous and authentic use of language by the learners because in tasks like story retelling, the demands of using specialized vocabulary and sequencing tense might put higher pressure on L2 learners’ cognitive processing (Préfontaine & Kormos, 2015). Additionally, in Tavakoli (2016), it is not clear whether the dialogic discussion tasks included authentic topics that ESL learners can easily relate to. One sample discussion topic mentioned in Tavakoli (2016) was “watching a movie at home or in the cinema”, which may not elicit spontaneous discussion from someone who does not watch or like movies. Moreover, in Michel et al. (2007) and Michel (2011), the monologic and dialogic tasks were on the same topic, which was likely to elicit spontaneous, authentic discussion (e.g., giving advice on buying a cell phone). However, Michel et al. (2007) and Michel (2011) did not examine how the CAF measures of the monologic and dialogic tasks were related to the L2 learners’ oral proficiency.

Therefore, although existing studies found significant differences in CAF measures between monologic and dialogic tasks (Ferrari, 2012; Michel, 2011; Tavakoli, 2016), there is a lack of investigations into whether CAF measures of monologic versus dialogic tasks are differentially related to L2 oral proficiency. Such investigations would have implications about the role of speaking task types (monologic versus dialogic) on linguistic predictors of L2 oral proficiency (that taps into the structural as well as the pragmatic aspects of oral production). Furthermore, several studies examined development in CAF measures over time (e.g., Ferrari, 2012; Tonkyn, 2012; Vercellotti, 2017). However, these studies did not examine development in L2 oral proficiency over time, and it is not clear whether CAF measures of oral performances predict longitudinal development in L2 oral proficiency. Such an investigation would offer important theoretical and pedagogical implications about linguistic features related to oral proficiency development.

In addition to the linguistic features, ESL speakers' individual difference (ID) variables might also be related to their attainment of proficiency in the L2 (Dörnyei, 2005) because ID variables refer to personal characteristics that everybody has but in varying degrees (Dörnyei, 2005). Among the ID variables, WM and aptitude have been widely investigated in relation to their effects on L2 oral skills. Previous studies found WM and aptitude to be significant predictors of linguistic features of L2 oral production (e.g., Kormos & Sáfár, 2008; Nielson, 2014; Granena, 2018). However, there is lack of a clear picture on how different components of WM and aptitude are related to L2 oral proficiency over time. Hence, the dissertation focuses on these two ID variables (WM and aptitude) as predictors of L2 oral proficiency.

2.5 Individual cognitive differences: WM

In cognitive psychology, WM refers to the systems of temporary maintenance and manipulation of information (Baddeley, 2012). The multicomponent model of WM (Baddeley & Hitch,1974; Baddeley, 1983,1986, 2000, 2003) is best-known because of its extensive use in research on higher-level cognition including both first language (L1) and L2 oral performances (Juffs & Harrington, 2011; Kormos & Sáfár, 2008). The multicomponent model of WM was divided into three components by Baddeley and Hitch (1974): central executive or executive working memory, henceforth, EWM (Baddeley, 2003; Juffs & Harrington, 2011) and two temporary storage systems. One is related to speech and sound: phonological loop that handles phonological memory, henceforth, PM (Baddeley, 2003, 2012; Juffs & Harrington, 2011), and the other is related to visuo-spatial aspects (visuo-spatial sketchpad) (Baddeley, 2012; Baddeley & Hitch, 1974). In L2 acquisition research, EWM and PM have been emphasized.

PM and EWM are argued to have fundamental and distinctive effects on L2 vocabulary and morphosyntactic development (Martin & N. Ellis, 2012; Wen, 2015). Whereas PM has been found to be important for spoken language development among children (Juffs, & Harrington, 2011), EWM may affect complex cognitive processes during L2 subskills learning (Wen, 2015). N. Ellis (2005) also argued that different components of WM are variedly related to different aspects of language learning. For example, PM is related to the memory of form and the ability to retain phonological information, whereas the EWM, measured by complex span tests, is associated with "explicit learning and the analysis of the language that is temporarily represented in the phonological loop or episodic buffer as well as in consciously created construction" (N. Ellis, 2005, p. 339). Due to such "overlapping involvements of the different components" of WM in different tasks, the present study included measures of both PM and EWM to get a more balanced estimate of the WM effect (N. Ellis, 2005, p. 339).

2.5.1 Phonological loop: The controller of PM

Phonological loop is a temporary verbal-acoustic storage system that is necessary for immediate retention of verbal or digital elements. Phonological loop consists of a "a brief store together with a means of maintaining information by vocal or subvocal rehearsal" (Baddeley, 2012, p.7). Thus, the phonological loop can be broken down into two sub-parts: a temporary storage component, which holds memory only for seconds unless that memory is rehearsed by a second component: a sub-vocal rehearsal system (articulatory component) (Baddeley, 2003). The rehearsal system maintains information and registers visual information in the store if the items can be named (Baddeley, 2003). The effect of phonological loop is on the storage of information related to order. The strength of this component is that it provides "temporary sequential storage, using a process that is rapid and requires minimal attention" (Baddeley, 2012, p. 12). In the multicomponent model, the phonological loop controls PM (Baddeley, 2003, 2012; Juffs & Harrington, 2011). A person's PM can be measured by simple word or digit span tests in which participants first hear a series of words/digits and then are asked to repeat those words/digits as accurately as they can. Hence, these tests tap into only phonological storage (Juffs & Harrington, 2011).

2.5.2 Central executive or EWM

Central executive or EWM is the most complex component of WM system (Baddeley, 2001, 2003, 2012). Its functions include the ability to divide attention between two targets or stimuli, interact with long-term memory, and manage the shifts between task performance and the retrieval processes that are necessary for task completion (Baddeley, 2012). More importantly, the EWM controls the attention that is necessary for maintaining focus and ignoring distracting information that might interfere with task completion (Juffs & Harrington, 2011). EWM processes are one of the main factors that determine individual differences in working memory span (Baddeley, 2003). Complex span tests (such as operation span tests) measure the ability to store information while doing additional processing tasks (Juffs, & Harrington, 2011; Linck et al., 2014). Thus, complex span tests tap into both the processing and storage functions of EWM (Baddeley, 2003; Juffs & Harrington, 2011). Complex span tests have also been successful in predicting achievement in complex cognitive tasks such as reading and reasoning (Baddeley, 2000).

In Baddeley and Hitch’s (1974) initial model, EWM was assumed to be an attentional system with no storage capacity (Baddeley & Logie, 1999). However, in Baddeley (2000), the fourth component of the WM model, "episodic buffer" was added. The episodic buffer functions as a temporary storage system that combines information from the sub-systems (phonological loop and visuospatial sketchpad) with those from the long-term memory and forms a basis for conscious awareness (Baddeley, 2003). The current multicomponent model of WM (Baddeley, 2000) is shown in the Figure 2.1.

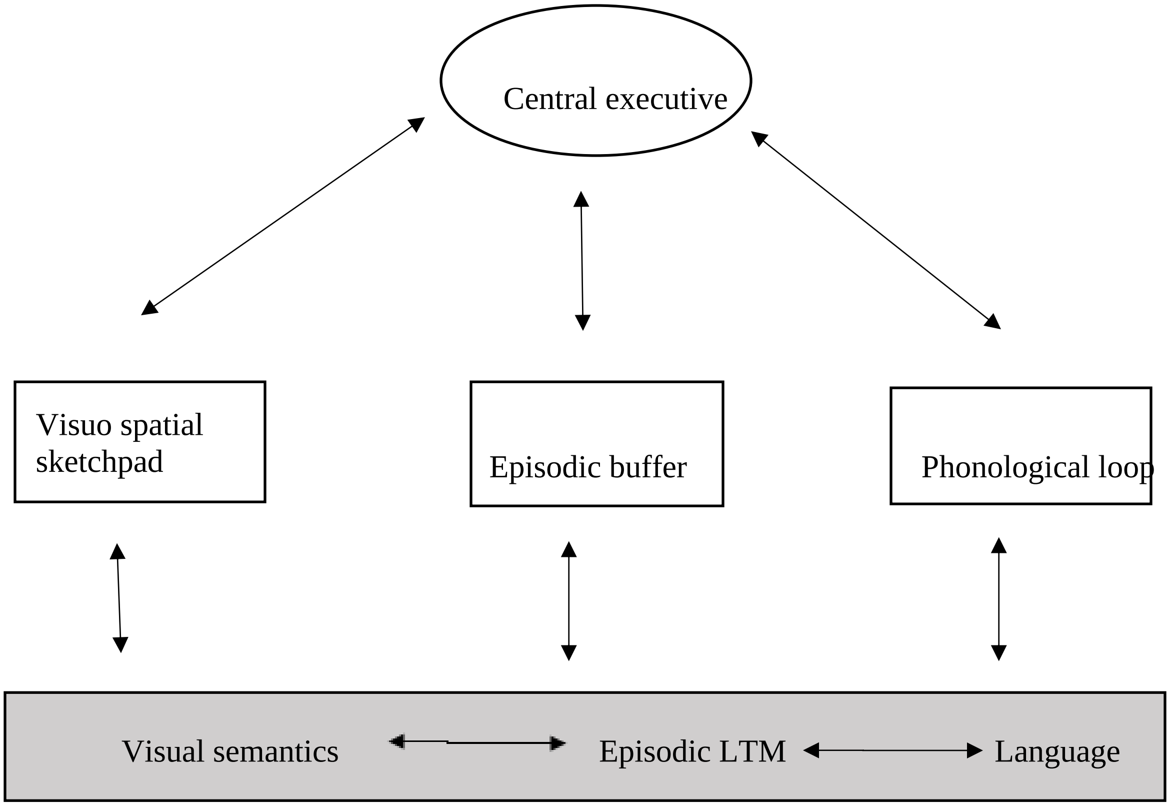

Figure 2.1 The Multicomponent Working Memory Model by Baddeley (2000, p. 421)

In Figure 2.1, the central executive, as the controller, is related to the other components (episodic buffer, visuospatial sketchpad and phonological loop), as indicated by the arrows, which reflect its role in coordinating resources between the temporary sub-systems (Wen, 2016a). The shaded area in the figure indicates the cognitive abilities that can gather long-term memory of linguistic and semantic knowledge (Baddeley, 2000). On the contrary, the components in the unshaded areas are argued to be "fluid" abilities i.e., temporary storage and attention (Baddeley, 2000, p. 418).

2.5.3 PM and L2 oral performance

Previous studies found that PM scores had significant correlations with the development in L2 oral fluency (O'Brien et al., 2006; O'Brien et al., 2007), comprehensibility, vocabulary, and syntax (Payne & Ross, 2005; Payne & Whitney, 2002). Additionally, some studies found significant effect of PM for beginner level learners. For example, Kormos and Sáfár (2008) and Révész (2012) found that for pre-intermediate or beginner level ESL learners, their nonword repetition test scores were significantly correlated with their scores in fluency and vocabulary (Kormos & Sáfár, 2008) and in past progressive construction (Révész, 2012). On the contrary, for participants of varied proficiency levels in Mizera (2006), PM was not significantly related to L2 fluency measures. Additionally, in Wen (2016b), above-intermediate level participants' PM scores were not correlated with any of the oral performance measures. Hence, for novice L2 learners, PM might be a more important bridge to oral fluency than for L2 learners of higher proficiency levels (Temple, 1997) although more studies are needed to verify such assumptions. Moreover, the available studies mostly focused on linguistic features of oral production (e.g., fluency), not on oral proficiency. More empirical investigations are needed to clarify the role of PM in the development of oral proficiency for L2 learners of varied proficiency levels.

Similar to the PM, there are also mixed findings about the effects of EWM on L2 oral performances. Such mixed findings might be partially due to the lack of uniformity in the way EWM has been measured in previous studies. For measuring EWM, some studies used verbal EWM tests (tests administered in the participants' L1 or in their L2), whereas some other studies used non-verbal span tests (that did not require any language processing in the processing component of the tests).

2.5.4 EWM and L2 oral performance

Some studies that used complex span tests in the participants' L2 found significant relationship between EWM capacity and CAF measures of L2 oral productions. For example, Mota (2003) used a L2 speaking span test with advanced level ESL learners and found significant relationship between the EWM and L2 fluency measures in monologic oral tasks.

However, scores in an L2 WM span test might be affected by the test-takers' L2 proficiency (Gass & Lee, 2011). Additionally, the findings of the studies that used WM span tests in the participants' L1 suggest that EWM might be more strongly related to L2 speakers' accuracy and fluency of oral performances than to syntactic complexity. For example, Ahmadian (2012) found positive correlations between intermediate level ESL learners' L1 listening span scores and their accuracy and fluency in an oral narration task although there was no such correlation between the span scores and the syntactic complexity measures. Similarly, Gilabert and Muñoz (2010) found significant correlations between both high and low proficiency learners' L1 reading span scores and measures of fluency and lexical variety in an L2 oral narration task, and only for the high proficiency participants, there was a moderate correlation between the span scores and the lexical complexity measure. Thus, learners' L2 proficiency might mediate the relationship between their EWM and L2 oral performances.

Furthermore, compared to the number of studies that used verbal or domain-specific span tests, there are only a few studies on L2 oral performance that employed domain-general or non-verbal complex span tests. One reason may be that mismatch between the content of complex span tests and the type of the language skills in focus may result in low correlations between them (Engle et al., 1999). For example, Kormos & Trebits (2011) did not find any strong correlation between ESL learners' EWM scores, measured by a backward digit span test, and the CAF measures of monologic oral narration of pictures. Similarly, in Mizera (2006), the scores of a math span test were not significantly correlated with Spanish as L2 learners' oral fluency scores in a similar oral performance task. However, there are also positive findings in this regard. For example, beginner level ESL learners in Kormos and Sáfár (2008) with high backward digit span scores had significantly higher gains on the accuracy and vocabulary parts of the oral test containing both monologic picture-description and dialogic problem-solving tasks. Similarly, Kim et al. (2015) found that EWM, measured by a domain-general running span test, was a significant predictor of English question development in interactive L2 oral productions tasks. Likewise, Nielson (2014) found that in the fluency measure (pruned speech rate) of an oral narrative task, ESL learners with higher EWM capacity, measured in a spatial span test, performed significantly better than those with lower EWM capacity.

Therefore, in literature, there are mixed findings about the relationships between EWM and different features of L2 oral production. Moreover, most studies examining the relationships between PM or EWM and L2 oral skills focused on L2 oral production features (e.g., CAF measures). Hence, there is a lack of investigations into whether PM and EWM are related to L2 oral proficiency that subsumes not only L2 speakers' knowledge of L2 forms but also their ability to use those forms appropriately in communicative contexts (Hulstijn, 2011). Additionally, as PM is related to the storage of information and EWM, to complex cognitive processes in L2 learning (Wen, 2015), PM and EWM might also be predictive of development in L2 oral proficiency over time. However, so far, there has been a lack of scholarly interests in this area.

Furthermore, in research literature, WM has often been incorporated as a component of aptitude (DeKeyser & Koeth, 2011; Skehan, 2012; Linck et al., 2013) although the meta-analysis of Li (2016) suggests the need for further empirical investigations in this regard. Aptitude is a cognitive variable in SLA, and it refers to any personal characteristic that is "important to achieving a learning goal including affective (feelings and emotions) and conative (goal setting and determination) processes such as anxiety and motivation as well as cognitive abilities such as analytic ability and memory" (Li, 2016, p. 803). It has been argued that all the components of aptitude (e.g., phonetic coding, language analytic ability, and memory) converge in WM (Miyake & Friedman, 1998). However, according to Li (2016), there needs to be more empirical investigations into how WM is related to aptitude. Whereas a good number of studies investigated how aptitude is related to L2 learning, only a few studies examined how aptitude is related to other cognitive variables such as WM (Li, 2016). The findings of the meta-analysis in Li (2016) showed significant and consistent correlations between aptitude and EWM but weak and nonsignificant correlations between aptitude and PM. Linck et al. (2014) also found stronger correlations between complex span tests measuring EWM and L2 achievement than between simple span tests measuring PM and L2 outcome. Thus, it is tempting to surmise that EWM is more likely to be a component of aptitude than PM (Li, 2016; Linck et al., 2014). However, whereas L2 achievement was treated as a composite construct in Linck et al. (2014), WM and aptitude might be differentially related to different aspects of L2 learning (for example, learning of grammar, vocabulary, development of distinct L2 skills, such as speaking, listening, reading, or writing) (Li, 2016). Therefore, more empirical research is needed to explore the relationships between the components of WM (EWM and PM) and aptitude regarding their effects on a specific aspect of L2 skill such as L2 speaking (Li, 2016). Because of the lack of a clear empirical evidence that WM and aptitude tap into the same underlying construct (Li, 2016), aptitude is treated in the dissertation as a distinct construct from WM.

2.6 Individual cognitive differences: Language aptitude

Aptitude is a cognitive variable in SLA and refers to any personal characteristic important in learning process including cognitive abilities such as analytic ability (Li, 2016, p. 803). Aptitude is related to "the ease and speed with which one learns a foreign language in comparison with peers during a certain period" (Li, 2016, p. 804). In Carroll’s (1981) view, aptitude is related to learners' readiness to learn a language, and it facilitates learning in formal instructional settings where learners make conscious effort to learn a foreign language. Four basic abilities for language learning were postulated by Carroll (1981): phonetic coding (the ability to analyze unfamiliar sound for retention), grammatical sensitivity (the ability to understand the functions of words in sentences), inductive learning (the ability to generalize and induce rules), and rote learning (the ability to associate verbal materials), and all these abilities combinedly underlie the construct of language aptitude (Sparks et al., 2011).Thus, L2 aptitude is componential where each component taps into a distinct language skill, and overall, the construct of aptitude has been highly predictive of L2 learning (Li, 2016; Sparks et al., 2011).

In recent research, a distinction has been made between cognitive aptitudes for explicit and implicit learning (Doughty et al., 2010; Granena, 2016, 2018; Linck et al., 2013). Language learning, as a part of human cognitive system, may be affected by individual predispositions for information processing in particular ways (Granena, 2016). Likewise, cognitive aptitudes for explicit and implicit learning were found to be related to two distinct categories of cognitive style: rational-analytical and experiential-intuitive, respectively (Granena, 2016). In Granena (2016), a rational-analytical style, that refers to the dependence on logic and analysis while processing information, was found to be related to explicit learning ability (Granena, 2016). On the contrary, an experiential-intuitive cognitive style, that refers to a tendency to depend on intuition and holistic thinking while processing information, was found to be significantly related to implicit learning ability (Granena, 2016). Thus, these two aspects of aptitude, explicit and implicit, suggest distinct language learning abilities. Explicit aptitude refers to conscious, controlled, and analytical processing of information, whereas implicit aptitude indicates intuitive, holistic, and automatic information processing (Granena, 2018).

Traditional aptitude tests mostly tap into conscious or explicit L2 learning abilities (Li, 2016). A widely used test battery of traditional aptitude is the MLAT (Modern Language Aptitude Test) (Carroll, 1990) which has strong predictive validity based on large samples of data collected from varied levels of L2 proficiency (Li, 2016). Recently developed LLAMA test (Meara, 2005) is modeled on the MLAT and is based on Carroll's perspectives on the nature of aptitude. LLAMA has four components: LLAMA B (a vocabulary learning test measuring the ability to learn novel words), LLAMA D (a sound recognition test that evaluates the ability to recognize sound sequences), LLAMA E (a sound-symbol association test measuring the ability to form new sound-symbol connections), and LLAMA F (a grammatical inferencing test measuring the ability to infer the rules of a novel language). The LLAMA test components "involve forming associations consciously and intentionally and working out relations in data sets," and hence, those are likely to "draw on explicit learning processes" (Granena, 2016, p. 583). However, LLAMA D may be an exception because it does not have a study phase unlike the other components, and it also does not require the use of analytical skills (Granena, 2013). Furthermore, in a principal component analysis in Granena (2013), both LLAMA D and a measure of implicit learning ability, serial reaction time (SRT) test (Kaufman et al. 2010), loaded under the same factor. Advanced language learning is related to practice in the target language environment for long-term that potentially involves implicit learning processes (DeKeyser, 2009). Therefore, DeKeyser (2009) argued for the importance of implicit aptitude or implicit learning ability.

In this regard, a recent development in aptitude test battery, Hi-LAB by Linck et al. (2013), included measures of implicit learning such as an SRT test that taps into implicit cognitive processes and was not measured in MLAT or in any traditional aptitude test batteries

(Granena, 2018). Additionally, in a recent study, Granena (2018) administered four LLAMA tests (LLAMA B, LLAMA D, LLAMA E, LLAMA F) and four tests belonging to the Hi-LAB: Paired Associates test, Letter Span test, SRT Test, and Available Long-Term Memory (ALTM) Synonym test on a population of 135 intermediate level college learners of L2 Spanish. In an exploratory factor analysis, Granena (2018) found that three LLAMA tests (LLAMA B, LLAMA E, and LLAMA F) and two Hi-LAB tests (Paired Associates and Letter Span) loaded under the same factor explaining 28.02% of the total variance, and that factor was labeled as explicit aptitude because those five tests measured explicit cognitive processes. On the contrary, the other two factors (explaining 17.05% and 14.55% of additional variance) had loadings from the tests (ALTM Synonym, LLAMA D, and SRT Test, respectively) that were argued to be related to implicit aptitude because they measured implicit cognitive processes (Granena, 2018). Granena (2018) further distinguished between implicit memory ability and implicit learning ability. Granena (2018) proposed ALTM Synonym and LLAMA D to be measuring implicit memory ability because these two tests involve retrieval of information. In addition, Granena (2018) proposed SRT test to be measuring implicit learning ability that is related to encoding of input. SRT test is based on assessing test-takers' sequence learning ability, which is one aspect of cognitive aptitude and is relevant for implicit language learning and processing (Granena, 2013). Suzuki and DeKeyser (2015) also used the SRT test as a measure of aptitude for implicit learning and found that the participants' SRT scores were significantly and positively related to their L2 implicit knowledge. Therefore, explicit and implicit aptitude might be distinct aspects of the aptitude construct and might have varying effects on L2 acquisition (Granena, 2016, 2018, 2019). While previous research, examining the relationships between aptitude and L2 oral skills, focused mostly on explicit aptitude (or conscious cognitive abilities), there has been a lack of research examining implicit aptitude in relation to L2 oral proficiency (Granena, 2019).

2.6.1 Aptitude and L2 oral performance

Among the studies examining the relationship between aptitude and L2 oral skills, Sparks et al. (2011) and Sparks et al. (1998) used ACTFL OPI for measuring L2 learners' (high-school level learners of French, German, and Spanish) oral proficiency. In both the studies, the participants' speaking proficiency was assessed on the following criteria: pronunciation, vocabulary, grammar, comprehensibility, and listening comprehension. Sparks et al. (2011) used MLAT to measure L2 aptitude and found that the measures of aptitude in addition to the participants' early L1 achievement, L1 cognitive ability, and L2 affective measures explained a considerable (76%) amount of variance in L2 oral and written proficiency. Similarly, Sparks et al. (1998) administered MLAT test on foreign language learners (of French, German, and Spanish) and found that the aptitude scores could significantly distinguish between learners with varied levels of oral proficiency.

Additionally, Granena (2018) investigated the extent to which the underlying constructs of the two aptitude test batteries, Hi-LAB and LLAMA, predicted CAF measures in a monologic oral picture description task performed by ESL learners. Granena (2018) found that the implicit memory ability, with significant loadings from LLAMA D and ALTM Synonym, predicted L2 oral fluency measured as pruned speech rate per minute. Thus, the participants with higher implicit memory ability had higher speech rate. The study also found that the learners with a broader productive vocabulary had higher implicit memory and implicit learning abilities. Overall, the findings of Granena (2013, 2018) showed that the construct of aptitude may encompass both explicit and implicit learning abilities, which might be differentially related to

L2 oral production features. Despite such empirical findings highlighting distinct components of the aptitude construct (explicit and implicit), previous research examining the relationships between aptitude and L2 oral skills mostly focused on explicit aptitude (Granena, 2019). In order to understand the effects of aptitude on L2 oral proficiency development, it is pertinent to examine how both explicit and implicit aptitude are related to proficiency in L2 speaking (Granena, 2019; Li, 2016).

2.7 Motivation for the present study

Based on the discussions above, several research gaps became apparent. Available SLA studies examining CAF measures and L2 oral proficiency mostly focused on monologic speech (e.g., De Jong et al.,2012b; Iwashita et al., 2008; Révész et al., 2016). Moreover, existing longitudinal studies on L2 oral production mostly examined the development in CAF features over time (e.g., Tonkyn, 2012; Vercellotti, 2017). However, there is not yet a clear picture of how CAF-based predictors of L2 oral proficiency vary over time or depending on oral task type (monologic versus dialogic). Such an investigation would offer insights not only into the importance of different task types (monologic versus dialogic) in determining linguistic predictors of ESL speakers' oral proficiency but also into the CAF variables related to L2 oral proficiency development over time. Individual differences in cognitive abilities (e.g., EWM, explicit aptitude) may also "lead to increasingly differentiated L2 speech production by learners on complex versions of tasks high in their reasoning demands" (Robinson, 2005, p. 58). Previous studies found significant effects of WM and aptitude on linguistic features (e.g. CAF measures) of L2 oral performances (e.g. Ahmadian, 2012; Fortkamp, 1999; Gilabert & Muñoz, 2010; Granena, 2018; Kormos & Sáfár 2008; Niwa, 2000). However, few studies examined whether different components of ESL speakers' WM (including both EWM and PM) and aptitude (including both explicit and implicit aptitude) are significantly related to their oral proficiency over time. Examining the effects of L2 speakers' cognitive variables (e.g., aptitude) in longitudinal research can offer insights into how the memory or learning ability is related to the development of proficient L2 performance over time (Skehan, 2016).

Moreover, in SLA research, there needs to be more focus on analyzing the processes of learning (DeKeyser, 2012), for example, understanding how L2 learning outcome is influenced by the interaction between objective features of language use and the subjective learner-related variables (Housen et al., 2019). DeKeyser (2012) also argued that one way of examining learning processes that are hard to observe is "to infer them from the way individual difference variables interact with linguistic and contextual variables" (p. 189). A linguistic variable (e.g., clausal complexity measure) may interact with an individual difference variable (e.g., explicit aptitude or EWM) in their effects on an outcome measure (e.g., L2 oral proficiency) because the linguistic measure might require a mental process that is facilitated (or hampered) by the individual difference variable (DeKeyser, 2012). For example, mean length of clause might be a significant predictor of L2 oral proficiency only for ESL speakers with higher explicit aptitude. Examining such interactions may not only indicate the importance of the internal WM/aptitude variables, the external linguistic variables, and their combined impact on the outcome measure (e.g., oral proficiency), but it can also offer insights into the process that links them (DeKeyser, 2012). DeKeyser (2012) discussed three sets of possible interactions involving aptitude (aptitudes x treatments[4], aptitudes x linguistic structures, and age x aptitudes) that can offer rich insights into efficient learning processes but are underrepresented in empirical research. Theoretically, examining the interactions between individual difference variables (such as

[4] By treatment, DeKeyser (2012) referred to "any kind of educational intervention at any level of generality, such as curriculum design, teaching method, content presentation, or practice activity" (p. 190).

aptitude) and linguistic structures can be informative of the processes underlying efficient L2 speech production. Practically, studying such interactions may allow for more fine-tuned and generalizable predictions of oral proficiency that can facilitate matching students with appropriate learning and practice activities (DeKeyser, 2012). Hence, it warrants investigation whether ESL speakers' WM or aptitude abilities interact with CAF measures of their oral production in their combined effects on L2 oral proficiency. However, there has been a lack of scholarly investigations into such interaction effects on L2 oral proficiency.

In response to these gaps in previous research, the goals of the dissertation study are threefold: (1) to investigate whether the relationships between CAF measures and L2 oral proficiency vary depending on task-type (monologic and dialogic) and over a period of eight months, (2) to examine whether ESL speakers' PM, EWM, explicit aptitude, and implicit aptitude predict any variation in their L2 oral proficiency over time, and (3) to investigate whether the relationships between ESL speakers' CAF measures and L2 oral proficiency are mediated by their WM and aptitude abilities. The study collects L2 oral performance data longitudinally using both monologic and dialogic oral tasks at three different periods (time one/ two/ three) over eight months. The study also expands the scope of data analysis in the current L2 oral production literature by using an NLP tool for measuring lexical sophistication.

2.7.1 Research questions

The dissertation has three main research questions, each of which has specific sub-questions. These are described below.

Research question 1: this question examines how CAF measures of monologic and dialogic oral tasks are related to L2 oral proficiency over time. Below are the specific sub-questions:

1a. Do the relationships between the CAF measures and L2 oral proficiency scores vary depending on task-type (monologic/ dialogic)?

1b. Do the relationships between the CAF measures of monologic and dialogic task performance and L2 oral proficiency scores change over time (time one/two/three)?

Research question 2: this question examines the relationships between WM and aptitude measures and L2 oral proficiency. Below are the specific sub-questions:

2a. Do ESL speakers' WM and aptitude measures predict their oral proficiency scores?

2b. Do the relationships between the WM and aptitude measures and L2 oral proficiency change over time (time one/two/three)?

Research question 3: this question examines the interactions between the WM/aptitude measures and the CAF indices in their effects on L2 oral proficiency. Below is the specific sub-question.

3. Do the relationships between the CAF measures of monologic and dialogic tasks and L2 oral proficiency vary depending on the participants' WM or aptitude abilities?

3 METHODS

3.1 Participants

Data were collected from 60 ESL speakers who were enrolled in different non-degree and degree programs at a public urban university in the Southeastern region of the USA. In the first phase of data collection, 88 participants signed up for the study; 76 of them returned to complete the second phase, and 60 participants completed the third phase. Hence, the total number of participants who completed all the phases of the study is 60. Among them, 22 participants were non-matriculated ESL learners in the Intensive English Program (IEP). Four IEP participants were in the high beginner level, 13 were in the intermediate, and five were in the advanced level.

These levels were based on the participants' performances in an in-house placement test at their entrance into the IEP. The remaining participants were from different matriculated programs. In contrast to the non-matriculated IEP learners who have not yet attained enough English proficiency to get admission to a degree program, the matriculated participants were L2 speakers of English enrolled in undergraduate or graduate degree programs. Among the matriculated participants, 21 were in the ESL-credit program, 9 were in different undergraduate programs, and 8 were in various graduate programs. Those in the ESL-credit program were also enrolled in undergraduate or graduate courses, but they were still required to take ESL classes to improve their English proficiency. Except the participants in the ESL-credit program, other matriculated participants did not need to take any ESL classes. Table 3.1 presents the demographic information of the participants, and Table 3.2, their L1 backgrounds. Among the participants, 25 were male and 35 were female. Each participant was paid $45 compensation for their participation.

Table 3.1 Demographic Information of the Participants

Program level	Number of participants	Age		Length of previous L2 study (in year)		Length of stay in the L2 country (in year)	
		Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Non-matriculated participants							
IEP	22	25	6.3	6.9	8.65	1.3	1.72
Matriculated participants							
ESL credit	21	25	5.3	9.9	3.79	1.8	3
Undergraduate	9	23	3.89	9	3.89	3.5	2.24
Graduate	8	26	4.2	9.6	5.04	2.45	3.19
Total	60	24.95	5.41	8.6	6.34	1.9	2.54

Table 3.2 L1 Backgrounds of the Participants

First language	Number of participants
Chinese	23
Arabic	13
Spanish	5
Korean	4
Vietnamese	3
French	3
Tigrigna	1
Malagasy	1
Portuguese	1
Turkish	1
Japanese	1
Bengali	1
Hindi	1
Nepali	1
Russian	1
Total	60

3.2 Materials

3.2.1 Oral performance tasks: Monologic and dialogic

In the dissertation, data were collected from each participant at three different times over eight months, and at each time, each participant completed one monologic and one dialogic task. Previous studies reported significant effect of topic familiarity on linguistic features of L2 oral performances (e.g., Bei, 2010). Hence, to compare the participants' performances over time and between tasks, all the monologic and dialogic tasks were based on the same topic: "Deciding where to live." However, to prevent any practice effect, six different versions of those tasks were created: six monologic versions (monologic A, B, C, D, E, and F) and six corresponding dialogic versions (dialogic A, B, C, D, E, and F) (see Appendices A, B, C, D, E, and F). Although each of those six versions were on the same topic (of choosing a place to live), the prompts focused on different aspects of the topic (for example, living in the downtown versus suburb, living on campus versus off campus, living in a house versus in an apartment). This decision was made to ensure that all the participants perform the speaking tasks on the same overall topic but on different prompts to obviate any practice effect.

In each task, the participants were provided input on the topic of the task, the context (that they were international students at a university in the USA and that they were looking for a housing), the options of two living places (including pictures) that they had to choose from, the question prompt, planning time, and the expected duration of their speech. In all the tasks, the participants chose a place to live, which is a situation commonly faced by international students in the USA. Hence, the tasks were related to authentic, real-word activities (R.Ellis, 2003; Loewen & Isbel, 2017). Based on the task-design framework of R. Ellis (2003, 2012), the design variables of the monologic and dialogic tasks in the dissertation study are listed in Table 3.3:

Table 3.3 Description of the Tasks in R. Ellis's (2003, 2012) Framework

Task-Design Variables	Monologic (“Deciding where to live”)	Dialogic (“Deciding where to live”)
Focused/Unfocused (on a specific L2 form)	Unfocused	Unfocused
Input-providing/ Output prompting	Output prompting	Output prompting
Cognitive process	Reasoning	Reasoning
Open/close	Open	Open
Structured/Unstructured	Structured	Structured
Goal orientation	N/A	Convergent

In each monologic version, the participants were provided with the context that they were international students at a US university and that they were looking for a place to live. Each monologic task presented descriptions of two options of living. The participants had to choose one and give reasons for their choice. The participants were given one-minute planning time before they started talking (Crowther, 2018), and the expected duration of their speech was mentioned as one to two minutes.

Each corresponding dialogic task was completed by pairs of participants, who were presented with similar context that they were international students in the USA and that they were looking for a housing to share. In each dialogic task, the participants were given the same two options of living as in the corresponding monologic task, and they were asked to choose one option to share with their respective partners. The input for each dialogic task was the same as that of the corresponding monologic task except that the dialogic input asked the participants to discuss their choices and the reasons behind those with their partners and come to an agreement about their choice of living. In the dialogic task, the participants received the same planning time, which was one minute, and the expected duration of their discussion was mentioned as two to four minutes.

3.2.2 Measures of oral proficiency

As the oral proficiency measures, the current study focuses on the TOEFL iBT speaking test and the communicative adequacy of monologic and dialogic oral tasks because these tests measure distinct aspects of the oral proficiency construct (Hulstijn, 2011). The rubric of the TOEFL iBT speaking test taps into the accuracy of syntactic forms, vocabulary, pronunciation features, and cohesive development of ideas (“TOEFL iBT Test,” 2014). Although the TOEFL iBT speaking rubric overlaps with the communicative adequacy rubric in terms of cohesion and comprehensibility, the communicative adequacy rubric, unlike the TOEFL speaking rubric, does not have any descriptor related to the use of language (e.g., complex structure and vocabulary). The rubric of communicative adequacy measures speakers’ ability to use L2 appropriately and comprehensibly for fulfilling communicative purposes (irrespective of the accuracy or complexity of their language) (Pallotti, 2009). Thus, these measures, in combination, correspond to the essential criteria of Hulstijn's (2011) definition of proficiency: knowledge of language forms and socially situated use of language.

3.2.3 Measures of oral proficiency: TOEFL iBT speaking test

In this dissertation study, one of the measures of oral proficiency was the participants' TOEFL iBT speaking test scores. Each participant at each time of data collection took one authentic TOEFL iBT speaking test. Three different TOEFL iBT speaking tests (Test A, Test B, Test C), previously administered by the ETS, were used for this purpose (ETS, 2016). 'Test A' is included in Appendix G as a sample. Those three tests were counterbalanced among the participants. TOEFL independent and integrated speaking rubrics (available at the following link, https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf) were used for rating the participants' responses.

TOEFL iBT independent and integrated rubrics are scaled from 0 (no attempt at speaking or unrelated response) to 4. In the rubrics, each speech is holistically evaluated based on the overall delivery of message (pronunciation and intonation features), language use (grammar and vocabulary), and development of topic (development of relevant and cohesive ideas) ("TOEFL iBT Test," 2014). A speech sample with a low score of 1 is characterized by the following: largely unintelligible speech, limited content, and/or minimal connection to the task ("TOEFL iBT Test," 2014). On the contrary, a sample with the highest score of 4 is characterized by the following: highly intelligible speech, high degree of automaticity in using basic and complex structures, and/or well-developed and coherent response to the task ("TOEFL iBT Test," 2014). For each participant at each time, their TOEFL iBT speaking test score was the average of their scores in all the six tasks.

3.2.4 Measures of oral proficiency: Communicative adequacy

The present study included communicative adequacy as a proficiency measure because it taps into functional effectiveness of language use, which is one aspect of the oral proficiency construct **(**Hulstijn, 2011). The monologic and dialogic oral performance data were used to measure the communicative effectiveness of the participants' speech. Hence, each participant at each time of data collection had two communicative adequacy scores: one for monologic and one for dialogic speech.

The rubric of Kuiken and Vedder (2018) (see Appendix H) was used for measuring the communicative adequacy of monologic speech. Kuiken and Vedder's (2018) rubric has four subscales, each measuring one of the four components of the communicative adequacy construct: content, task requirement, comprehensibility, and coherence and cohesion. Each sub-scale contains descriptors that are independent of linguistic features and are objective and countable (Kuiken & Vedder, 2018). The participants were scored on each of the four sub-scales (content, task requirement, comprehensibility, and coherence and cohesion) separately, and each participant was assigned an overall mean score. For example, if a participant received three for content, three for task requirements, four for comprehensibility, and four for coherence and cohesion, then their overall communicative adequacy score for monologic speech would be 3.5 (summation of the scores 3+3+4+4=14 divided by the number of categories [4]) (Crowther, 2018).

The rubric of Kuiken and Vedder (2018) was developed for rating monologic speech. Therefore, for rating the dialogic performances in the present study, the sub-scale of "communicative skills/strategies" from the "paired assessment rating rubric" of Ockey (2011) (see Appendix I) was added to the rubric of Kuiken and Vedder (2018). The paired assessment rubric, established in Ockey (2009,2011) and used in Leaper and Riazi (2014) and Crowther (2018), was developed as a measure of oral group performance at Kanda University of International Studies (Japan). Hence, in the present study, the rubric for rating dialogic oral performances contains five sub-scales: content, task requirement, comprehensibility, coherence and cohesion, and communicative skills/strategies. The descriptors of the category of communicative skills/strategies in Ockey (2011) are related to the participants' nature of interactions, their level of confidence, and awareness of conversational features (Ockey, 2011).

For the communicative adequacy of dialogic speech, each participant from a pair received a separate score. The communicative adequacy score of each participant consisted of an overall mean score based on their score in each sub-scale (similar to the communicative adequacy monologic scores). For example, if on a dialogic task, a participant received four for content, four for task requirements, four for comprehensibility, three for coherence and cohesion, and three for communication skills/strategies, then their overall communicative adequacy score for dialogic speech would be 3.6 (summation of the scores 4+4+4+3+3= 18 divided by the number of categories [5]) (Crowther, 2018).

3.2.5 Measures of WM: PM

As PM (phonological memory) and EWM (executive working memory) are argued to have fundamental and distinctive effects on L2 acquisition (Martin & N. Ellis, 2012; Wen, 2015, 2016a, 2016b)**,** the present study employed measures of both PM and EWM (N. Ellis, 2005). A forward digit span test, one of the widely used verbal memory tests, was used to measure the participants' PM (Kim et al., 2016; Olsthoorn et al., 2014; Révész, 2012). In this test, participants heard a series of random digits (one digit per second) and repeated the digits in the presented order (Kim et al., 2016). Their oral repetition responses were audio-recorded. The participants heard the sequences of digits in spans whose lengths ranged from three to nine digits, and those spans were presented in order of increasing length (Kim et al., 2016). Each span contained four lists of numbers. Previous research showed that participants' familiarity with the language of the digits might be a confounding variable in forward digit span test results (Thorn & Gathercole, 2001). Therefore, to avoid the confounding effect of L2 proficiency on the participants' performances in the digit span test, multiple parallel forms of a forward digit span test were developed in the participants' first languages (L1). Therefore, each participant in the present study performed the digit span test in their L1. For creating the digit span tests, the researcher first audio-recorded the numbers in different languages, and then the audio-editing software Audacity (“Audacity,” 2019) was used to create the tests.

To score the digit span tests, the current study adopted the partial scoring method where one point was assigned to a correct recall of a number at the correct position (Kim et al., 2016). As the total number of digits that the participants were asked to repeat was 168, the maximum score possible in this test was also 168 (Kim et al., 2016).

3.2.6 Measures of WM: EWM

Due to the confounding effect of test-takers’ L2 proficiency on their performances in EWM span tests administered in their L2 and the infeasibility of administering a complex span test in participants’ L1 in an ESL context (Gass & Lee, 2011), the present study used non-verbal span tests that do not involve any language processing in the processing component of the tests. Moreover, in addition to the EWM capacity, a WM span test might measure other factors irrelevant to the EWM construct such as the ability to solve math problems (in case of an operation span test) (Foster et al., 2014). Hence, a WM span test might contain "variance" from both the EWM capacity and the task itself (Foster et al., 2014, p. 2). Therefore, Foster et al.

(2014) argued that "researchers should use multiple indicators to create either a composite or factor score" of the EWM construct that "consists of the variance shared between two or more complex span tasks" (p. 2). Therefore, the dissertation used two shortened versions of complex span tests designed by Foster et al. (2014): one block of operation span (OSpan) test and one block of symmetry span (SymSpan) test. This combination of tests accounted for 40.1% variance in the fluid intelligence and WM factor in Foster et al. (2014). These tests also accounted for 78.5% of the total 51% variance in the fluid intelligence and WM factor explained by a full model combination (three blocks from each of OSpan, SymSpan, and rotation span tests) in Foster et al. (2014).

The OSpan test used simple math problems as the distractor task and letters as the items to-be-remembered (Kane et al., 2004; Foster et al., 2014). Participants first solved a simple math equation, then saw a letter, and then solved another math problem, and then saw another letter. For each trial, this math-letter sequence was repeated from three to seven times, each time with an unpredictable length (Foster et al., 2014). After each trial of math-letter sequence, participants had to recall, in order, the preceding letters. Scores were calculated by the summation of the number of letters accurately recalled in the correct order, which is also known as the partial score (Foster et al., 2014; Turner & Engle, 1989). As a single block of the OSpan test of Foster et al. (2014) contains 25 to-be-remembered items, the total score possible in the OSpan test in the present study was 25. The program running the test only outputted the total score.

The SymSpan (Kane et al., 2004) task has a method similar to the OSpan with a few main differences. First, the distractor task was to judge whether a displayed shape is symmetrical along with its vertical axis. Secondly, the items to-be-remembered were locations of red squares in a 4x4 grid of possible locations. Thirdly, the number of symmetry-location pairs ranged from two to five times per trial. The scores were calculated by adding the number of red square locations correctly recalled in the accurate order (the partial scoring method) (Foster et al., 2012). As a single block of the SymSpan test of Foster et al. (2014) contains 14 to-be-remembered items in sequence, the total possible score in this test in the present study was 14. The program running the test only outputted the total score.

As mentioned before, Foster et al. (2014) argued that for a reliable measurement of a psychological construct such as EWM, SLA studies need to create a composite or factor score from multiple WM span tests that included varied types of processing tasks (e.g., solving math problem, judging symmetry of shapes etc.). Hence, the dissertation study used the summation of the OSpan and the SymSpan scores as operationalization of the EWM.

3.2.7 Measures of aptitude: LLAMA tests

As aptitude measures, the current study used LLAMA B, LLAMA E, LLAMA F, LLAMA D (Meara, 2005), and a probabilistic SRT test (Kaufman et al., 2010; Suzuki & DeKeyser, 2015). As explained in the chapter 2, the set of LLAMA tests was developed by Meara (2005), and these tests focus on test-takers' conscious and explicit learning ability. LLAMA has four sections, LLAMA B, LLAMA D, LLAMA E, LLAMA F, which assess L2 speakers' ability to learn new vocabulary, recognize sounds, associate sounds with symbols, and infer logical rules, respectively. For each LLAMA test, the score was automatically calculated out of a total of 100 by the software that runs the test, and the program outputted only the total score for each test.

LLAMA B is a vocabulary learning task that measures test-takers' ability to learn large amounts of vocabulary (real words from a Central American language) in a short time (Meara, 2005). In the task phase, test-takers got 120 seconds to learn the names of as many of the 20 objects as they can. In the testing phase, the name of an object showed up on screen and the participants identified the picture of that object from the name.

LLAMA D measures whether a test-taker can recognize short stretches of spoken language that they were exposed to. In the task phase, 10 words from an unfamiliar language were played. In the test phase, they heard those words in addition to other words that they had not heard before. The scores were based on recognizing the words that were repeated in the test. The participants lost points for making a wrong choice, and they received points for every right choice (Meara, 2005).

LLAMA E is a sound-symbol association test. LLAMA E has a set of 22 recorded syllables with a transliteration of those syllables in an unfamiliar alphabet. The test-takers' responsibility is to find out the relationship between the sounds they hear and the writing system. In the task phase of this test (two minutes long), the participants clicked small buttons to hear short sound files, one by one, and the text on each button tell them how that specific sound is written in an unfamiliar alphabet (Meara, 2005). In the testing phase, the program played new sounds, one at a time, and simultaneously, it displayed two possible spellings for that word of which only one spelling was correct. The participants clicked on the correct spelling. They received points for every correct response and lost points for every wrong answer (Meara, 2005).

LLAMA F is a grammatical inferencing task in which the test-takers had 300 seconds to learn as much as they could about a new language. For each click of a button, they saw a picture and a sentence describing that picture. In the testing phase, the participants saw a picture and two sentences of which only one was grammatically correct. They chose the sentence that they thought was accurate. This test contained 20 test items in total (Meara, 2005).

3.2.8 Measures of aptitude: Probabilistic serial reaction time (SRT) test

Probabilistic SRT test, developed by Kaufman et al. (2010), indicates test-takers' ability of sequence learning and is argued to be a measure of implicit aptitude (Granena, 2013, 2018; Kaufman et al, 2010; Suzuki & DeKeyser, 2015; Willingham et al., 1989). For the dissertation study, a web version of the test was developed.

In the SRT test, the participants saw a stimulus at one of four locations on a computer screen. Their task was to press the corresponding key on the keyboard as fast and accurately as possible as they saw the stimulus on the screen. Four keys on computer keyboard, "V", "B", "N", and "M" corresponded to the four locations on the screen: "V" corresponding to the leftmost location, "M" corresponding to the rightmost location, and "B" and "N" corresponded to the middle left and middle right locations respectively. The participants were asked to place their second and third fingers (index and middle) of each hand on the four keys before starting the test so that they could respond as fast as possible.

During the task, the participants saw the stimulus in the four locations on the computer screen in a repeating sequence (training sequence) 85% of the time, which was intermixed with an alternate sequence (control sequence) 15% of the time. More specifically, Sequence A (1-2-1-4-3-2-4-1-3-4-2-3), the probable or the training sequence, occurred with a probability of 0.85, whereas Sequence B (3-2-3-4-1-2-4-3-1-4-2-1), the improbable or control sequence, occurred with a probability of 0.15 (Kaufman et al., 2010). Because of the probabilistic nature of the SRT task, it is hard to learn the sequence explicitly (Suzuki & Dekeyser, 2015).

At the beginning, the participants did a practice block where the training and control sequences may occur with equal probability (Kaufman et al., 2010). After the practice block, the participants did eight training blocks in which the stimulus followed Sequence A 85% of the trials and Sequence B, 15% of the trials. The participants completed 120 trials in a block, thus, 960 trials in total.

For assessing learning for each participant in the probabilistic SRT test, first, the error responses were deleted as well as the outlier responses (1.6% of the data) that were more than three standard deviations away from the mean in each block for each participant (Kaufman et al., 2010). The amount of learning was indicated by the difference between the average response time (RT) in the training condition and the average RT in the control condition from the third to the last block. The RTs in the training condition are likely to be faster than those in the control condition from block three to block eight, and the learning effect is not likely to surface on blocks one and two, which were not considered in the analysis (Kaufman et al., 2010; Suzuki & Dekeyser, 2015). For calculating the SRT scores, instead of measuring exact difference in the RT, the test assessed whether the participants showed a learning effect as large as the learning effect evident in the sample across blocks three to eight (Granena, 2013; Suzuki & Dekeyser, 2015; Kaufman et al., 2010). For each participant in each block, it was assessed whether their mean RT for probable trials was less than the difference between their mean RT for improbable trials and their standard deviation for RT on improbable trials[5] (Kaufman et al., 2010). If it was less, the participants received 1, and if it was not, they received 0. The total score for each participant was calculated by summing up their scores across the last six blocks (from blocks three to eight) with the minimum total score 0 and the maximum total score 6. The scores were calculated by using a SPSS scoring script collected from the first author of Kaufman et al. (2010) (through personal communication).

[5] The standard deviation for RT on improbable trials was also multiplied by the average difference in RT between the conditions (probable and improbable) across the blocks three to eight (Kaufman et al., 2010)

The reliability of the SRT test, calculated by split-halves with Spearman-Brown, was 0.39. This reliability index is low but is very close to the split-half reliability scores reported in previous studies using the same test, for example, 0.42 reported in Suzuki and Dekeyser (2015) and 0.44 reported in both Granena (2016) and Kaufman et al. (2010). Behavioral tests like SRT, that measure implicit learning ability, often have relatively low reliabilities (Kaufman et al., 2010; Reber et al., 1991).

3.3 Data collection procedure

As shown in Table 3.4, data were collected at three different times over eight months: Time 1 (September'2019), Time 2 (December'2019-January'2020), and Time 3 (April-May'2020). Data for the first two phases of the study was collected in a controlled laboratory setting. Due to the worldwide COVID-19 pandemic in spring of 2020, data for the last phase was collected online by Zoom video calls keeping all the other data collection procedures consistent ("Zoom Video Communications", 2020). Thus, for the last phase of data collection, the procedure was the same as in the previous two phases with the only exception that the participants completed the tasks online (using Zoom video calls).

For recruiting participants, the researcher visited IEP oral communication classes of all levels and ESL credit speaking/listening classes during the first week in the Fall'2019 semester. During the visits, the researcher explained the purpose of the study and provided the students the informed consent forms. The researcher also provided the interested students a sign-up sheet with available laboratory time-slots for the first phase of data collection (time one, September'2019) and requested that two students sign-up for each time slot. While scheduling the participants, the researcher also scheduled any two participants for each available time slot.

Table 3.4 The Procedure of Data Collection

Time	Month	Total approx. time required	Activities Completed (2 students appear together at each time)	Approx. time required for each task/test
Time one (Lab meeting)	September'2019	61 minutes	Dialogic task Monologic task TOEFL speaking test 2 EWM tests	4 minutes 2 minutes 20 minutes 35 minutes
Time two (Lab meeting)	December'2019- January'2020	51 minutes	Dialogic task Monologic task TOEFL speaking test 4 LLAMA tests	4 minutes 2 minutes 20 minutes 25 minutes
Time three (Zoom meeting)	April-May 2020	51 Minutes	Dialogic task Monologic task TOEFL speaking test PM test SRT test	4 minutes 2 minutes 20 minutes 10 minutes 15 minutes

At time one, the participants came to the lab (two students at the same time) at each scheduled time, and first, they did the dialogic task. After the dialogic task, two participants were placed in two separate rooms in the lab where they completed the other tasks. The order of the monologic and dialogic tasks of different versions was counterbalanced among the participants to avoid any ordering effect (Mackey & Gass, 2005). For counterbalancing the oral tasks, each participant was assigned to one of the six groups shown in Table 3.5.

Table 3.5 Counterbalancing the Monologic and Dialogic Tasks

	Time One	Time Two	Time Three
Group 1 (n=8)[6]	Monologic A Dialogic B	Dialogic D Monologic C	Monologic E Dialogic F
Group 2 (n=9)	Dialogic C Monologic B	Monologic D Dialogic E	Dialogic A Monologic F
Group 3 (n=10)	Monologic C Dialogic D	Dialogic F Monologic E	Monologic A Dialogic B
Group 4 (n=10)	Dialogic E Monologic D	Monologic F Dialogic A	Dialogic C Monologic B
Group 5 (n=14)	Monologic E Dialogic F	Dialogic B Monologic A	Monologic C Dialogic D
Group 6 (n=9)	Dialogic A Monologic F	Monologic B Dialogic C	Dialogic E Monologic D

At time one, after completing the oral tasks, the participants did the TOEFL iBT speaking tests and the two EWM tests (operation span and symmetry span). The order in which they did the TOEFL test and the EWM tests was changed for each participant. For counterbalancing the three TOEFL iBT speaking tests among the participants over three times of data collection, the design presented in Table 3.6 was followed. The order of the two EWM tests was also changed for each participant. The total time required at time one was about 60 minutes.

[6] At time one, data was collected from 88 participants, but 28 of them dropped out by time three. Hence, the number of participants in each group was not even.

Table 3.6 Counterbalancing the TOEFL Speaking Tests

Groups of Participants	Time 1	Time 2	Time 3
Group 1 (n=10)	Test A	Test B	Test C
Group 2 (n=10)	Test A	Test C	Test B
Group 3 (n=9)	Test B	Test A	Test C
Group 4 (n=13)	Test B	Test C	Test A
Group 5 (n=12)	Test C	Test A	Test B
Group 6 (n=6)	Test C	Test B	Test A

Before beginning the data collection for the second phase of the study (time two), the researcher contacted the participants again by email and text messages to schedule them for the available laboratory slots. Similar to time one, at time two, two participants were scheduled at the same time, and they first did their dialogic tasks followed by the monologic tasks maintaining the counterbalancing order of Table 3.5. Then they did the TOEFL iBT speaking tests (following the counterbalancing order of Table 3.6) and the LLAMA tests. The order in which they did the TOEFL test and the LLAMA tests was changed for each participant. The order in which they did the four LLAMA tests was also changed for each participant. At Time two, the total time required was about 50 minutes. At time one and time two, for audio-recording the participants' monologic oral performances, "SpyCenter Micro Voice Recorder MP3" was used, and for video recording the dialogic performances, "Sony Handycam CX405 Flash Memory Full HD Camcorder" was used.

Before beginning the data collection for time three, the researcher again contacted the participants (by emails and text-messages) for scheduling. For each available time slot, the researcher scheduled two participants. Because of the COVID-19 pandemic, the participants were scheduled for Zoom video meetings. For the participants, the researcher prepared an easy instruction sheet on how to download and install the Zoom software and how to join a Zoom meeting. When the researcher contacted the participants for scheduling for time three, the instruction sheet was sent to each of them by email. However, each participant was already familiar with Zoom video calls because most of them were attending their online classes and/or meetings on Zoom. Thus, although zoom meetings were new to the current study, the participants were already familiar with attending video meetings on Zoom. Few minutes prior to the meeting with each pair of participants, the researcher shared with them a Google Drive folder with the monologic and dialogic speaking tasks and the audio files for the PM test (digit span test).

At the beginning of each scheduled meeting, the researcher connected both the participants on Zoom. Then the researcher told them what dialogic task file to open (following the counterbalancing order of Table 3.5) from the shared Google Drive folder. The participants did the dialogic speaking task together. After the dialogic task, one participant was disconnected from the Zoom call while the other participant completed all the other scheduled tasks: the monologic task (following the counterbalancing order of Table 3.5), the TOEFL speaking test (following the counterbalancing order of Table 3.6), the PM test, and the SRT test. Once the participant completed all the scheduled tasks and disconnected, the researcher connected again with the other participant who was disconnected before to let him/her complete the tasks. The order in which they did the TOEFL test, the PM test, and the SRT test was changed for each participant. For administering the TOEFL tests, the researcher ran the tests on her own laptop (with the Zoom recording mode on) and shared the screen with the participants with both the "share computer sound" and "optimize screen-sharing for video clip" options checked so the participants could hear the audio of the shared screens. The participants' responses to the TOEFL test-prompts were recorded on Zoom. For the PM tests, the participants played their respective digit span test file (audio-recorded in their L1) from the shared Google Drive folder, and the researcher recorded their repetitions of the digits on Zoom. For the web version of the SRT test, the researcher shared a link with the participants in Zoom chat box. The participants clicked the link and started the test. The test was accompanied with clear instructions, and the researcher also stayed in the video meetings throughout the entire time to answer any question they might have. The participants' responses to the SRT test were saved by the program that ran the test. After the zoom meeting with each pair of participants, the researcher saved all the audio and video recordings in her laptop computer ("Lenovo ThinkPad T480").

3.4 Data coding

3.4.1 Preparing the oral performance data for analysis

The recorded monologic and dialogic oral performance data were saved in the researcher's laptop computer. For transcribing the monologic data, a professional transcription service was used. The researcher double checked the accuracy of each transcribed monologic file. All the transcriptions were done verbatim including false starts, repetitions, and self-corrections. Before transcribing the dialogic data, each dialogic video was converted to an audio file using an online app ("123apps", 2020). For transcribing the dialogic data, PRAAT (version 6.0.37, Boersma & Weenink, 2018) was used, and the converted audio files were uploaded to PRAAT (in .wav format) for transcription. The researcher did the transcription of each dialogic file (transcriptions done verbatim including all the false starts, repetitions, and self-corrections). For each participant in a dialogic pair, a separate transcribed text-grid file was created on PRAAT. Another researcher, experienced in transcribing ESL speech, double checked the accuracy of 10% of the transcribed dialogic data and reported 99% accuracy. From the transcribed monologic and dialogic data sets, greetings (e.g., “hello”, “good morning”) from the beginnings and any closing “thank you” from the ends were deleted. Thus, only the participants’ responses to the task prompts were used for the CAF analysis. An overview of the constructs (including oral proficiency, the oral production features [i.e., CAF], and individual difference variables) in the current study and their respective operationalizations is presented in Table 3.7.

Table 3.7 The Constructs in the Present Study and their Operationalizations

	Constructs	Operationalizations
Features of L2 Oral Production	Syntactic Complexity	Mean length of AS-unit Subordination measure Coordination measure Mean length of clause Frequency of wh-clauses
	Lexical Sophistication	MRC familiarity all words MRC meaningfulness all words COCA spoken frequency all words
	Accuracy	Number of error-free AS-units
	Fluency	Mean length of pauses Mean length of fluent runs Phonation-time ratio Articulation rate False starts per 100 words Text-length
L2 Oral Proficiency	Oral Proficiency	TOEFL iBT speaking test score Communicative adequacy score of monologic speech Communicative adequacy score of dialogic speech
Individual Difference Variables	EWM	Operation Span test Symmetry Span test
	PM	Forward digit span test
	Aptitude	LLAMA B LLAMA E LLAMA F LLAMA D SRT Test

In the dissertation, monologic and dialogic oral performances were analyzed in terms of complexity at both the levels of lexis and syntax that are crucial aspects of L2 speech performance **(**Skehan, 2009). The oral performances were also analyzed for utterance fluency that often determines the overall oral proficiency of a speaker (Huensch & Tracy-Ventura, 2017). The oral production features (i.e., the CAF constructs) and their respective operationalizations are explained below.

3.4.2 Analysis of oral production features: Syntactic complexity

The dissertation study included both general and specific measures of syntactic complexity (Bulté & Housen, 2012; Robinson & N. Ellis, 2008)**.** As general complexity measures, four types of indices were used: measures of subordination, coordination, phrasal complexity, and overall sentence complexity (Bulté & Housen, 2012; Norris & Ortega, 2009). As the measure of overall sentence complexity, *Mean Length of Analysis of Speech (AS) unit* was calculated: total number of words divided by the total number of AS-unit (Foster et al., 2000). While counting the total number of words for calculating *Mean Length of AS-unit*, the present study did not consider false starts (an utterance, which is started and then abandoned altogether or reformulated in some way), repetitions (repeating previously produced speech), and self-corrections (an error identified by the speaker either during or immediately following production and corrected) (Foster et al., 2000). The AS-units were identified based on the definition of Foster et al. (2000), "an AS-unit is a single speaker's utterance consisting of an independent clause or sub-clausal unit, together with any subordinate clause(s) associated with either" (p. 365). Following the level one analysis of oral performances in Foster et al. (2000), the dissertation used all the transcribed data for identifying the AS units in the monologic speech. For identifying the AS units of the dialogic speech, following Foster et al. (2000), the dissertation study excluded turns consisting of only one word minor utterances (e.g., "yes," "no," "okay," "right") whose inclusion could "distort the perception of the nature of the performance" (Foster et al., 2000, p.370).

The *Subordination Measure* was operationalized as the number of clauses per AS-unit (total number of clauses divided by the total number of AS-unit). As per the definition in Foster et al. (2000), a subordinate clause was operationalized as consisting of at least "a finite or non-finite verb element plus at least one other clause element (Subject, Object, Complement or Adverbial)" (p.366).

Table 3.8 and Table 3.9 show examples of AS units with single clause and multiple clauses from monologic and dialogic task performances, respectively.

Table 3.8 Example of AS-units from a Monologic Speech in the Dissertation

<table>
<tr><th>Examples</th><th>Annotation</th></tr>
<tr><td>|I prefer :: to live in residential area |</td><td>1 AS-unit 2 clauses</td></tr>
<tr><td>|It is more cheaper|</td><td>1 AS unit 1 clause</td></tr>
<tr><td>|There is {a lot of} a lot of advantages|</td><td>1 AS unit, 1 clause, 1 repetition</td></tr>
<tr><td>|That is :: why I prefer :: to live away from the downtown.|</td><td>1 AS-unit, 3 clauses</td></tr>
</table>

Note. The notations of AS-unit analysis (Foster et al., 2001) are used to indicate AS-units (enclosed in two upright slashes, | |), clauses (divided by two double colons, ::), false starts and repetitions(within curly brackets, {}).

Table 3.9 Example of AS-units from a Dialogic Speech in the Dissertation

<table>
<tr><th>Examples from person A's dialogic speech</th><th>Annotation</th></tr>
<tr><td>Person A: |as for the bigger one it seems luxury|
|So I think ::it is very expensive.|
|I cannot afford it|</td><td>3 AS units 4 clauses</td></tr>
<tr><td>Person B:….</td><td></td></tr>
<tr><td>Person A: yes</td><td>('yes' excluded from the analysis)</td></tr>
<tr><td>Person B:….</td><td></td></tr>
<tr><td>Person A: okay that was very good explanation</td><td>1 AS unit 1 clause</td></tr>
</table>

Note. The table only includes person A's speech (person's B's speech is omitted). The notations of AS-unit analysis (Foster et al., 2001) are used to indicate AS-units (enclosed in two upright slashes, | |), clauses (divided by two double colons, ::), false starts and repetitions(within curly brackets, {}).

The dissertation study operationalized *Coordination Measure* as the number of coordinated clauses per AS-unit (total number of coordinated clauses divided by the total number of AS-unit) (Bulté & Housen, 2012). Each clause that started with any of the coordinating conjunctions ("and," "but," "or," "so," "for," "nor" or "yet") was coded as one coordinated clause. The current study also calculated *Mean Length of Clause* that captures syntactic complexity sub-clausally at the phrasal level (Norris & Ortega, 2009). Each participant's *Mean Length of Clause* was calculated by dividing the total number of words (excluding false starts, repetitions, self-corrections) by the total number of clauses (Foster et al., 2000).

As the specific complexity measure, the dissertation study calculated *Frequency of Wh-Clauses (per 100 words)* (Bulté & Housen, 2012; Révész et al., 2016). This measure was selected because an examination of the data collected during the pilot administration of the tasks showed that while explaining reasons behind their choices of housing, ESL speakers used Wh-clauses (as complement) more frequently than other structures (e.g., to-infinitive, auxiliaries) that have been suggested in L2 literature (e.g., Bulté & Housen, 2012; Révész et al., 2016) as specific complexity measures. For *Frequency of Wh-Clauses (per 100 words),* all clauses initiated with wh-words were manually counted.

The coding for syntactic complexity were done by the researcher and a second coder coded randomly chosen 20% of the data. The agreement between the raters was 98%. All the disagreements were resolved through discussion. After the calculation of inter-rater reliability, the researcher coded the rest of the data.

3.4.3 Analysis of oral production features: Lexical sophistication

For measuring lexical sophistication, an NLP tool, TAALES version 2.2 (Kyle & Crossley, 2014) was used. The dissertation used word frequency scores calculated with the Corpus of Contemporary American English (COCA) spoken as the reference corpus (Davis, 2008-). The COCA spoken contains 104 million words, and it was created from the transcripts of unscripted conversations from about 150 radio and television shows collected between 1990 and 2015. For calculating the COCA spoken frequency scores, TAALES only considers the words in the target texts (e.g., dataset of the present study) that also appear in the appropriate database (in this case, COCA spoken). The dissertation considered COCA spoken frequency scores for all words (both content and function words). *COCA Spoken Frequency All Words* in TAALES is calculated by dividing the sum of the frequency scores for words in a text by the number of words in that text that received a frequency score (Kyle & Crossley, 2014, p. 766).

The dissertation study also focused on psycholinguistic word information indices in TAALES that reflect the depth of L2 speakers' word-knowledge (Salsbury et al., 2011). TAALES calculates the psycholinguistic word information indices from the Medical Research Council (MRC) psycholinguistic database (Coltheart, 1981), and the dissertation considered indices for all words (both content and function words). The dissertation included the following psycholinguistic word information indices from TAALES: *MRC Familiarity All Words* and *MRC Meaningfulness All Words*. Familiarity indicates how familiar to adults a word is, and meaningfulness indicates how related a word is to other words (Salsbury et al., 2011). For calculating the indices, TAALES divides the sum of familiarity or meaningfulness scores of a text by the number of words in that text that received a familiarity or meaningfulness score, respectively (Kyle &Crossley, 2014). *MRC Familiarity All Words* was a significant predictor of L2 speaking proficiency and *MRC Meaningfulness Content Words*, a significant predictor of L2 lexical proficiency in Kyle and Crossley (2014).

3.4.4 Analysis of oral production features: Accuracy

Based on previous studies (e.g., Foster & Wigglesworth, 2016) that argued for using global measures of accuracy based on a syntactic unit (e.g., AS-unit or clause) in L2 oral performance, the dissertation operationalized accuracy as the *Number of Error-Free AS-units* (per 100 words) (Ferrari, 2012; Tonkyn, 2012). For identifying the error-free AS-units, data were first segmented into AS-units. Then each AS-unit was manually coded for errors in both grammar and lexis (Révész et al., 2016). All kinds of grammatical errors were considered (e.g., errors in article use, preposition, subject-verb-agreement, tense, sentence structure etc.). If an AS-unit contained an inappropriate or inaccurate use of lexis, it was also considered as an error. All the coding for accuracy were done by the researcher, and a second coder coded randomly chosen 20% of the data. The exact agreement between the raters was 97%. All the disagreements were resolved through discussion.

3.4.5 Analysis of oral production features: Fluency

Oral fluency was operationalized as breakdown fluency (measured by *Mean Length of Pauses, Phonation-Time Ratio,* and *Mean Length of Fluent Runs*[7]), speed fluency (measured by *Articulation Rate*), and repair Fluency (measured by *False Starts per 100 Words*) (De Jong & Perfetti, 2011; De Jong et al., 2012a; Révész et al., 2016). In the dialogic tasks, the unclaimed between-turn pauses were excluded from the analysis (Tavakoli, 2016). Additionally, in the dialogic recordings, there were a few instances of overlapping interactions. Each of those overlaps was considered as belonging to the speaking time of both the speakers involved and hence, were counted in the fluency measures of both the speakers (Tavakoli, 2016).

For counting the *Mean Length of Pauses*, total duration of pauses (in seconds) was divided by the number of pauses. *Phonation-Time Ratio* was calculated as the ratio of the total length of time spent speaking (in seconds) and the total utterance time (in seconds) including the pauses (Prefontaine & Kormos, 2015, p.99). Thus, for calculating *Phonation-Time Ratio,* the total time

[7] Mean length of run was considered as a measure of breakdown fluency in De Jong et al. (2012) and as a measure of speed fluency in Tavakoli and Skehan (2005).

filled with speech (excluding all pauses) was divided by the total time spent speaking (time filled with speech + silent and filled pauses) (De Jong & Perfetti, 2011). Furthermore, *Mean Length of Fluent Runs* was the mean number of syllables produced between pauses, and this measure was calculated by dividing the total number of syllables by the total number of runs (De Jong & Perfetti, 2011). Tavakoli (2016) operationalized *Mean Length of Fluent Runs* and *Phonation-Time Ratio* as composite measures of fluency because they tap simultaneously into rates of speech and pauses, thus, blending the speed and flow of speech (Tavakoli, 2016, p. 138). Skehan (2014) suggested that while operationalizing fluency, such composite measures should be considered.

Additionally, *Articulation Rate* (a measure of speed fluency) was calculated as the total number of syllables divided by the length of speaking time excluding the pauses (De Jong et al., 2012a; De Jong & Perfetti, 2011). Text-length (number of words) was also included as a fluency measure to control for the confounding effect of the amount of speech produced by the participants (see the “Statistical Analysis” section for more details). Furthermore, repair fluency was operationalized as the *Number of False Starts per 100 Words* (Révész et al., 2016). A false start was identified as “an utterance which is begun and then either abandoned altogether or reformulated in some way” (Foster et al., 2000, p. 368). The dataset did not include many examples of repetitions and self-corrections that have also been used in literature as indices of repair fluency. Hence, as the repair fluency measure, the dissertation focused on *Number of False Starts per 100 Words.*

The software program PRAAT version 6.0.37 (Boersma & Weenink, 2018) was used to measure the duration of pauses, the duration of soundings, the number of runs, and the number of pauses. These measures were used to calculate all the fluency indices. First, each sound file was uploaded to PRAAT in the .wav format. Then each file was annotated using the “To Text Grid (silences)" feature of PRAAT. The "minimum silent interval duration" (the minimum duration of silence detected by the program) was set to 250 milliseconds (De Jong & Bosker, 2013; Kahng, 2014). Nonverbal fillers such as "uh," "um," "mmm," "aaa" were counted as pauses (De Jong & Perfetti, 2011, Vercellotti, 2017). In each annotated Text Grid file, the "silent" and "sounding" boundaries were manually checked for accuracy by both listening to the recording and examining the spectrogram and waveform. Furthermore, an online syllable counting tool ("Syllable Count," 2018) was used to count the syllables of the transcribed audio files. The tool uses an US English dictionary (containing 240,364 words) and a syllable-counting algorithm to count syllables ("Syllable Count," 2018). The accuracy of syllable-counting was also manually checked by the researcher for 10% of the data, and the rate of accuracy was 99%.

All the fluency coding was done by the researcher. A second rater coded 20% of the randomly selected data, and the exact agreement between the two coders was 98%. Any disagreement was resolved through discussion.

3.4.6 Oral proficiency ratings: Ratings of TOEFL iBT speaking tests

Each TOEFL iBT speech sample was scored by two expert raters who were doctoral students in applied linguistics with previous experience in rating ESL speech. The raters also went through a training. At first, the raters together rated 3 audio-recorded responses (of actual test-takers taking the test) to TOEFL iBT speaking tests. After discussion, they independently rated about 15 of those responses, compared ratings with each other (*r* = 0.85), and discussed any differences. Then they independently rated the TOEFL test responses from the present study. Table 3.10 reports the Pearson correlation and Cohen's Kappa between the raters' scores for each time of data collection. In cases where they diverged by only one point, the middle point was assigned to the participants. When they diverged by more than one point (below 2% of the entire ratings), they discussed their disagreements and came to a decision about the ratings. Each participant's TOEFL iBT speaking test score at each time of data collection was the average score of all the six tasks.

Table 3.10 Interrater Reliability Scores between the Two Raters of the TOEFL iBT Speaking Test

	Pearson's r	Cohen's Kappa
Time One	0.85	0.81
Time Two	0.88	0.85
Time Three	0.88	0.70

3.4.7 Oral proficiency ratings: Ratings of communicative adequacy

Two native English speakers (university undergraduate students majoring in applied linguistics and in their third year) were recruited as raters of communicative adequacy. Each monologic and dialogic recording was rated by those two raters (Crowther, 2018; Révész et al., 2016). Neither of the raters had previous experience of teaching ESL or rating ESL speech although as applied linguistics major, they might still have had the metalinguistic knowledge. Previous studies examining the communicative adequacy construct (De Jong et al., 2012b; Kuiken & Vedder, 2018) recruited such non-expert raters because a rater with previous experience of teaching ESL might find it hard to focus entirely on the communicative effectiveness of L2 speech ignoring grammatical errors. Furthermore, in Révész et al. (2016), there was only a little difference between the expert and non-expert raters in their communicative adequacy rating of ESL speech. For rating, the raters used the audio-recordings of the monologic speech and the video recordings of the dialogic performances (Crowther, 2018).

For rating the communicative adequacy, each rater received training. First, each rater took a few days to be familiar with the rubrics. Then during the training session, they collaboratively rated four speech samples (collected during the pilot administration of the tasks) representing two high performing (one monologic, one dialogic) and two low performing (one monologic, one dialogic) samples determined by the researcher. Then, each rater independently rated five additional speech samples, including both monologic and dialogic speeches. They discussed and resolved any differences in their ratings. After that, they independently rated 20 monologic oral performances from the present study with the correlation 0.88 (*r* = 0.88). Then the raters independently rated all the oral performances from the present study. Table 3.11 reports the inter-rater reliability scores (Pearson's r and Cohen's Kappa) for the communicative adequacy ratings of the monologic and dialogic tasks for each time of data collection. In a case the raters differed by one point while rating a sub-scale in the rubric, the middle point was assigned to the participant. In the few cases when the raters differed by more than one point, they discussed their differences and collaboratively came to a decision. This same procedure was followed for rating both the monologic and the dialogic tasks.

Table 3.11 Interrater Reliability Scores for the Communicative Adequacy Ratings

	Communicative Adequacy for Monologic Tasks	
	Pearson's r	Cohen's Kappa
Time One	0.76	0.70
Time Two	0.77	0.66
Time Three	0.73	0.70
	Communicative Adequacy for Dialogic Tasks	
	Pearson's r	Cohen's Kappa
Time One	0.84	0.76
Time Two	0.79	0.70
Time Three	0.76	0.71

3.5 Statistical analysis

To answer the research questions, linear mixed effect (LME) analyses were performed to account for the random variance associated with the participants at different time periods. The lme4 package (version 1.1-15) in the software program R (version 3.4.3) was used for the LME analyses (R Core Team, 2015). In all the LME models, participant was the random intercepts to account for the variance related to the participants. The participants' L1 background was added as a fixed factor in all the models to control for their individual differences in L1 that can have important influences on L2 oral skills (Crossley et al., 2018; Derwing & Munro, 2013; Ringbom & Jarvis, 2011). The participants' L1 background was operationalized as language distance scores (Adsera & Pytlikova, 2015; Crossley et al., 2018) reported in Chiswick and Miller (2005). Chiswick and Miller (2005) showed that other conditions of language learning (e.g., length of instruction) being equal, a lower score in learning a language by native English-speaking Americans indicates greater distance between that language and English. Hence, linguistic distance scores represent the distance between a selected language and English based on native English speakers' difficulties in learning that language. Each L1 is assigned a score from 1 to 3 where a score of 1 (e.g., Japanese) is farther from English than a score of 3 (e.g., Romanian). Additionally, as the participants were recruited from both non-matriculated IEP and the matriculated programs (ESL credit, undergraduate and graduate), "program level" (with two categorical levels: IEP and matriculated) was also included as a fixed factor in all the models to control for any variance explained by the participants' program affiliation. Moreover. in all the models that included CAF measures as the predictors, text-length was added as a fixed factor (predictor) to control for the potential confounds of the length of the participants' speech (Linck & Cunnings, 2015). The R function r.squared GLMM was used to calculate the effect sizes where the marginal r squared (R^2m) indicates the variance explained by the fixed factors, and the conditional r squared (R^2c) indicates the variance explained by both the fixed and random factors.

As the dissertation used multiple variables (TOEFL iBT speaking test, communicative adequacy for monologic tasks, and communicative adequacy for dialogic tasks) as operationalizations of the oral proficiency construct, a factor analysis was conducted to examine whether those three types of scores loaded under the same factor. For each of these oral proficiency measures, there were three sets of data in the dissertation because data were collected from each participant three times. Hence, a multiple factor analysis (MFA) was conducted, which is an extension of principal component analysis (PCA) (Bécue-Bertaut & Pagès, 2008). In an MFA, the influence of different sets of data is balanced (Bécue-Bertaut & Pagès, 2008). Additionally, to measure aptitude, which is a componential construct (Li, 2019), the present study used five tests: LLAMA B, LLAMA E, LLAMA F, LLAMA D, and the SRT. Hence, a PCA was conducted to examine whether there is any reduction in the dimensions of the aptitude construct. Although there is no definite sample size for factor analyses, in SLA studies that used a factor analytic approach, the median variable-to-participant ratio was 12 (Loewen & Gonulal, 2015). Hence, the sample size of the present study (n=60) is consistent with the usual sample size used in factor analyses in SLA literature for the MFA analysis conducted on the three oral proficiency variables (3 x 12=48) and also for the PCA analysis on the five aptitude variables (5x12=60). The FactoMineR package from the software program R was used for the MFA and PCA analyses (Husson et al., 2020).

Additionally, there are different opinions in literature regarding the optimal value of factor loadings (Loewen & Gonulal, 2015). For determining the significance of factor loadings, the present study adopted the criteria of Stevens (2009) who offered guidelines for evaluating factor loadings based on sample size. According to Stevens (2009), if sample size is smaller, there might be considerable opportunity for capitalization on chance in factor analysis and rotating, which might lead to higher amount of errors in factor loadings. Hence, Stevens (2009) argued that the significance of factor loadings should be determined based on sample sizes and proposed critical values of factor loadings (at alpha 0.01) for sample sizes ranging from 50 to

1000. Based on the critical values of Stevens (2009), for the sample size of the current study (n=60), the minimum value of significance for factor loadings was set at 0.61.

For selecting the appropriate CAF measures for the LME models for answering the research questions 1a and 1b, first the correlations between all the CAF measures and the oral proficiency score (dependent variable) were examined. Any variable with a correlation at or below 0.10 ($r <= 0.10$) and/or with a non-significant *p*-value ($p > 0.05$) were discarded for reporting a small effect size. Multicollinearity was also checked using the Variance Inflation Factors (VIF) scores (with the “vif ()” function in the “car” package of R, Levshina, 2015) and by checking the correlations between the fixed factors. Any fixed factor having a correlation with another fixed factor above 0.70 ($r > 0.70$) was discarded (Mostafa & Crossley, 2020). After checking for correlation and multicollinearity, the selected fixed factors were added one-by-one to a null LME model, and after adding each fixed factor, “anova” was used to compare each model with the previous one to examine any significant difference between the models (Levshina, 2015). The variables whose addition made a significant difference ($p < 0.05$) were selected to enter the final LME model. All the linguistic indices were transformed into z-scores to maintain the uniformity of scaling.

For answering the research question 1a (whether the relationships between the CAF measures and L2 oral proficiency scores vary depending on task-type [monologic/ dialogic]), two LME models were developed: model 1 for monologic data and model 2 for dialogic data. Monologic and dialogic tasks have different information processing demands, which might differentially affect the production of L2 speech (Robinson, 2005). Therefore, separate models were built for monologic and dialogic tasks for addressing the research question 1a. In both the models, selected CAF measures along with text-length, L1 distance, and program level were the fixed factors or predictors, and the oral proficiency score was the response (dependent) variable. Additionally, at each time, the participants responded to a different speaking prompt (on the same topic, “deciding where to live”). Hence, prompt was added (in both the models 1 and 2) as random intercept to explain the variance related to prompt[8]. Additionally, in the dialogic tasks, some participants talked to the same partner at each time of data collection, while some participants talked to a different partner at each time. To account for these variations related to participant pairing in the dialogic tasks, a variable labelled “pair combination” was created and included as random intercept in the model 2 (on dialogic data). To create this “pair combination” variable, each pair of participants at each time was assigned a number (starting from 1). For example, if a participant was paired up with the same partner at each time for the dialogic tasks, they were assigned to the same number for pair combination. However, if a participant was paired up with a different person at each time of data collection, then that participant was assigned a different number at each time for pair combination.

Thus, in both LME model 1 (for monologic data) and LME model 2 (for dialogic data), selected CAF measures (along with text-length, L1 distance, and program level) were included as the fixed factors, oral proficiency scores was the dependent variable, and participants and prompts were the random intercepts. Additionally, in the model 2 (for the dialogic data) “pair combination” was also included as the random intercept. Then, in each model, the non-significant predictors with the higher *p*-values were discarded one by one until the model was left with only the significant predictors, and the model with the significant predictors was reported.

[8] Different participants might respond differently to the within-subjects factor of prompt, and this random effect can be explained by including by-participant random slopes for prompt. However, as the model failed to converge when ‘prompt’ was added as a random slope (by participants), in the final model, ‘prompt’ was included as random intercept.

To answer the question 1b (whether the relationships between the CAF measures and the oral proficiency scores change over time [time one/two/three]), interactions were fitted in the LME models 1 and 2 between the CAF measures and time (one/two/three). For each of the LME models (model 1 for monologic data and model 2 for dialogic data), first interactions were fitted between all the CAF measures and time. Then, the non-significant interaction effects with the higher *p*-values were discarded one by one until the model was left with only the significant interactions. Likewise, the non-significant fixed effects with higher *p*-values were discarded one by one until the model was left with only the significant fixed effects. The tables with the significant fixed and interaction effects are reported.

To select the appropriate fixed factors for answering the research question 2a (whether L2 speakers' WM and aptitude measures predict their oral proficiency scores), first, the correlations of the WM and aptitude variables with the oral proficiency scores (response variable) were checked. The indices with correlation at or below 0.10 (r =< 0.10) and/or a non-significant *p*-value ($p > 0.05$) were discarded for reporting a small effect size. Multicollinearity was also checked, and any WM and aptitude measures with correlations with another measure above 0.70 ($r > 0.70$) was discarded. Then an LME model (model 3) was created with the selected WM and aptitude scores as the fixed factors or predictors and the oral proficiency scores as the dependent variable. L1 distance and program level were also included as predictors to control for differences in the participants' L1 backgrounds and program affiliations. Additionally, to answer the question 2b (whether the relationships between the WM and aptitude measures and the oral proficiency scores change over time [one/two/three]), interactions were fitted in the LME model 3 between the WM and aptitude measures and time (one/two/three) to examine if the WM and aptitude measures affected the oral proficiency scores over time.

For answering the research question 3 (whether the relationships between the CAF measures and the oral proficiency scores are mediated by the participants' WM and aptitude variables), two LME models (model 4 for monologic data and model 5 for dialogic data) were developed with the oral proficiency scores as the dependent variable. The information processing demands of different tasks (e.g., monologic versus dialogic) may variedly interact with L2 speakers' cognitive abilities (Carroll, 1990; Robinson, 2005b), which can affect the efficacy of the linguistic features used in those tasks. Hence, separate models were developed for monologic and dialogic tasks for addressing the research question 3. In the LME model 4 (for monologic data), participants and prompt were included as the random intercepts, and in the LME model 5 (for dialogic data), participants, prompt, and pair-combination were included as the random intercepts. Additionally, the CAF measures included in the models for answering the questions 1a and 1b and the WM and aptitude measures included in the models for answering the questions 2a and 2b were the predictors or fixed factors (along with text-length, L1 distance, and program level) in the LME models 4 (for monologic data) and 5 (for dialogic data) for answering the research question 3. These models (model 4 for monologic data and model 5 for dialogic data) fitted interactions between the CAF measures and the WM/aptitude scores to examine whether the effects of the CAF measures on the oral proficiency scores varied depending on the participants' WM/aptitude. For each of the LME models (model 4 and 5), first interactions were fitted between all the CAF measures and the WM/Aptitude variable. Then, the non-significant interaction effects with the higher *p*-values were discarded one by one until the model was left with only the significant interactions. Likewise, the non-significant fixed effects with the higher *p*-values were discarded one by one until the model was left with only the significant fixed effects. The tables with the significant fixed and interaction effects are reported. In all the LME models, a fixed factor was considered significant if the *p*-value was below 0.05.

4 RESULTS

The first set of research questions in the dissertation examines whether the relationships between the CAF measures and L2 oral proficiency vary depending on task-type and time. The second set of questions examines whether ESL speakers' WM and aptitude scores predict their L2 oral proficiency over time. The third research question examines whether the relationships between the CAF measures and the oral proficiency scores are mediated by the participants' WM and aptitude abilities. Prior to creating LME models, factor analyses were conducted to reduce the dimensions of the constructs which were measured using multiple tests. For example, an MFA was conducted to test whether the three variables (TOEFL iBT speaking, communicative adequacy for monologic tasks, and communicative adequacy for dialogic tasks) for measuring L2 oral proficiency loaded under the same factor. A PCA was also conducted to examine whether the tests used for measuring aptitude (LLAMA B, LLAMA D, LLAMA E, LLAMA F, and the serial reaction time [SRT] test) loaded under a smaller number of variables. In the beginning of the "Results" section, the output of these statistical tests used for dimension reduction (MFA and PCA) are reported before the results of the LME models. For each research question, first, the descriptive statistics of the predictor variables and the results of the procedures followed for variable selection are described. Then the results of the LME models are narrated to answer the research question.

4.1 Results of the MFA analysis

Table 4.1 presents the descriptive statistics of the oral proficiency measures that were collected at three different times over eight months.

Table 4.1 Descriptive Statistics of the Oral Proficiency Measures

	Time one				
	Scale	Mean	Std. Dev	Min	Max
TOEFL iBT	0-4	2.24	0.76	0.16	3.62
Communicative adequacy monologic	1-6	4.61	0.94	1	6
Communicative adequacy dialogic	1-6	4.92	0.82	1.8	6
	Time two				
TOEFL iBT	0-4	2.63	0.74	0.58	3.91
Communicative adequacy monologic	1-6	4.29	0.73	1.25	5.75
Communicative adequacy dialogic	1-6	4.63	0.70	1.8	5.65
	Time three				
TOEFL iBT	0-4	2.61	0.68	0.91	3.75
Communicative adequacy monologic	1-6	4.46	0.55	3.25	5.68
Communicative adequacy dialogic	1-6	4.46	0.62	1.9	5.6

Note. Std. Dev=Standard deviation; Min= Minimum; Max= Maximum

Before conducting the MFA, the Bartlett test was conducted to examine whether these three variables (TOEFL iBT speaking, communicative adequacy for monologic tasks, and communicative adequacy for dialogic tasks) tapping into the oral proficiency construct were correlated (Levshina, 2015). The *p*-value of the Bartlett test ($\chi^2 = 158.44$, $p < 0.001$) was well below the significance level indicating that the null hypothesis of zero correlation between the variables can be rejected. Table 4.2 shows the correlations among the three proficiency measures.

Table 4.2 Correlations between the Oral Proficiency Measures

	TOEFL iBT score	Communicative adequacy monologic	Communicative adequacy dialogic
TOEFL iBT score	1	0.59	0.52
Communicative adequacy monologic	0.59	1	0.57
Communicative adequacy dialogic	0.52	0.57	1

Table 4.3 shows the eigenvalues (the proportions of the total variance explained by each dimension or factor, Levshina, 2015) of the dimensions from the MFA output.

Table 4.3 The Eigenvalues of the Dimensions from the MFA Output

Dimensions (Factors)	Eigenvalue	Percentage of variance	Cumulative percentage of variance
Dimension 1	2.689	66.595	66.595
Dimension 2	0.321	7.957	74.552
Dimension 3	0.274	6.788	81.340
Dimension 4	0.214	5.292	86.632
Dimension 5	0.190	4.706	91.338
Dimension 6	0.140	3.468	94.806
Dimension 7	0.108	2.685	97.491
Dimension 8	0.059	1.467	98.958
Dimension 9	0.042	1.042	100.000

The higher are the correlations between a dimension and the variables, the higher is that dimension’s eigenvalue (Levshina, 2015). According to Kaiser criterion, which is commonly used, only those dimensions should be retained whose eigenvalues are greater than 1 (Levshina, 2015). Additionally, in Jolliffe’s criterion, all eigenvalues above 0.70 should be retained (Loewen & Gonulal, 2015). The present study adopts Kaiser criterion for selecting dimension or factor, and as can be seen in Table 4.3, only dimension 1 meets this criterion with eigenvalue greater than 1. This dimension also explains about 66% of the total variance. Moreover, as shown in Table 4.4, all the three proficiency variables had the highest and significant factor loadings (correlations between variables and a dimension or factor) under dimension 1.

Table 4.4 The Factor Loadings from the MFA Analysis

Groups	Dimension 1
TOEFL iBT score	0.935
Communicative adequacy (Monologic)	0.966
Communicative adequacy (Dialogic)	0.941

Based on the values proposed by Stevens (2009) for determining significant factor loadings, the loadings with absolute value above 0.61 were considered as significant for the sample size of the current study (n=60). Hence, all the three oral proficiency variables had significant factor loadings under dimension 1 (with the loadings ranging from 0.93 to 0.96). Previous research (e.g., DiStefano et al., 2009; Loewen & Gonulal, 2015) suggested the use of summation or average of each participant's scores on the variables that comprise a factor as the factor score. Hence, for each participant at each time, the dissertation study used the summation of the following three scores as the oral proficiency variable (the dependent variable) in the statistical models: average TOEFL iBT speaking score + average communicative adequacy monologic score + average communicative adequacy dialogic score. Table 4.5 shows the descriptive statistics of the oral proficiency scores that were used in the statistical models as the dependent variable.

Table 4.5 Descriptive Statistics of the Oral Proficiency Scores Used in the Statistical Analyses

	Mean	Std. dev	Min	Max
Time one	11.76	2.14	3.46	15.18
Time two	11.55	1.91	3.63	14.90
Time three	11.53	1.59	6.06	14.68
Overall	11.61	1.88	3.46	15.18

Note. Std. dev=Standard deviation, Min= Minimum, Max=Maximum

4.2 The PCA analysis on the aptitude variables

In the dissertation study, a total of five tests were used to measure the construct of aptitude. Table 4.6 shows the descriptive statistics of the aptitude test scores.

Table 4.6 Descriptive Statistics of the Aptitude Test Scores

	Mean	Std. Dev	Min	Max
LLAMA B	47.33	19.05	15	85
LLAMA D	25.41	12.99	0	60
LLAMA E	81.67	21.08	10	100
LLAMA F	46.17	25.31	0	90
SRT	1.78	1.37	0	6

Note. Std. Dev=Standard Deviation, Min=Minimum, Max=Maximum
The maximum possible score in each LLAMA test=100 and in the SRT test=6

As can be seen in Table 4.6, the participants' aptitude scores ranged from 0 (e.g., in LLAMA F) to 100 (e.g., in LLAMA E). The *p*-value (χ^2 = 24.208, *p*=0.007) of the Bartlett test was below the significance level (0.05) indicating that the null hypothesis of zero correlation between the variables could be rejected. Table 4.7 shows the eigenvalues of the dimensions from the PCA output.

Table 4.7 Eigenvalues of the Dimensions from the PCA Output

	Eigenvalue	Percentage of variance	Cumulative percentage of variance
Dimension 1	1.667	33.352	33.352
Dimension 2	1.094	21.881	55.234
Dimension 3	0.972	19.444	74.678
Dimension 4	0.815	16.318	90.996
Dimension 5	0.450	9.003	100.00

As can be seen in Table 4.7, only the dimensions 1 and 2 have eigenvalues above 1 that meet the Kaiser criterion (Levshina, 2015), and together they explain 55% of the total variance. Table 4.8 shows the factor loadings for each of these two dimensions:

Table 4.8 Factor Loadings of the Selected Dimensions from the PCA Analysis

	Dimension 1	Dimension 2
LLAMA E	**0.789**	
LLAMA F	**0.807**	
LLAMA B		-0.593
LLAMA D	0.519	
SRT	0.26	**0.807**

Note. Significant factor loadings are in bold font

As can be seen in Table 4.8, for the dimension 1, the factor loadings of only LLAMA E and LLAMA F met the criteria of significance (of minimum correlation 0.61, Stevens, 2009), and for the dimension 2, only SRT met this criterion. In the PCA conducted on 135 L2 learners' LLAMA test scores in Granena (2018), LLAMA E and LLAMA F also loaded under the same factor, which was named as explicit aptitude. LLAMA E and LLAMA F measure test-takers' analytical skills and explicit inductive learning ability (Granena, 2016, 2018, 2019). Hence, in the dissertation, for the factor score of the dimension 1, the average of LLAMA E and LLAMA F was calculated, and this dimension was named "*Explicit Aptitude*". Additionally, similar to Granena (2018), SRT in the present study is significantly loaded under a separate dimension. SRT has been used in the literature as a measure of implicit learning ability (e.g., Granena, 2013, 2018; Kaufman et al., 2010; Suzuki & DeKeyser, 2015), that is, "an individual's ability to learn a pattern or rule through simple exposure and without the intent to learn the pattern" (Granena, 2018, p. 17). Therefore, the output of the PCA conducted on the dataset of the dissertation shows two significant dimensions of the aptitude construct (which also support the results of a similar analysis in Granena, 2018): *Explicit Aptitude* (LLAMA E and LLAMA F; average=64, standard deviation=20) and SRT (Serial Reaction Time) scores. As LLAMA B and LLAMA D did not significantly load under any dimension, those are retained in the study as separate aptitude scores.

4.3 Research question 1a: Whether the relationships between the CAF measures and L2 oral proficiency scores varied depending on task-type (monologic/ dialogic)

Table 4.9 lists the descriptive statistics of all the CAF measures (including the text-length) for the monologic and dialogic tasks at each of the three time (one, two, three).

Table 4.9 Descriptive Statistics of the CAF Measures

Syntactic complexity: *Mean Length of Clause*						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	5.75	0.94	5.9	0.88	5.98	0.86
Dialogic	5.45	0.80	5.32	0.85	5.52	0.70
Syntactic complexity: *Subordination Measure*						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	1.95	0.52	2.36	0.70	2.59	0.72
Dialogic	1.44	0.30	1.80	0.37	1.89	0.40
Syntactic complexity: *Coordination Measure*						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	0.52	0.20	0.63	0.30	0.74	0.23
Dialogic	0.36	0.15	0.41	0.17	0.48	0.20
Syntactic complexity*: Mean Length of AS-Unit*						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	11.21	3.33	14	3.98	15.35	3.99
Dialogic	7.90	2.10	9.64	2.59	10.46	2.53
Syntactic Complexity: *Number of Wh-Clauses per 100 words*						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	1.20	1.08	1.11	1.03	1.06	0.88
Dialogic	1.30	1.17	1.48	1.09	1.28	0.91
Lexical Sophistication: *MRC Familiarity All Words*						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	602.31	4.12	601.40	3.84	601.45	3.62
Dialogic	601.73	3.68	602.71	3.90	602.62	3.61
Lexical Sophistication: *MRC Meaningfulness All Words*						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	369.29	12.46	367.57	11.54	366.82	14.29
Dialogic	370.18	12.54	374.73	14.4	369.08	11.77
Lexical Sophistication: *COCA Spoken Frequency All Words*						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	7988	1444	7787	1164	7884	1196
Dialogic	7327	2071	7791	1578	7934	1115
Accuracy: *Number of Error-Free AS-units*						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	5.72	2.26	3.84	1.84	3.68	1.75

Dialogic	7.19	2.46	6.61	2.41	5.79	2.23
Fluency: *Mean Length of Pauses*						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	0.83	0.26	0.78	0.20	0.75	0.14
Dialogic	0.69	0.19	0.64	0.13	0.67	0.12
Fluency: *Mean Length of Fluent Runs*						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	5.95	2.37	6.01	2.19	5.99	1.82
Dialogic	5.69	1.78	6.03	1.85	6.38	2.12
Fluency: *Phonation-Time Ratio*						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	0.63	0.10	0.65	0.10	0.65	0.08
Dialogic	0.72	0.10	0.74	0.08	0.72	0.07
Fluency: *Articulation Rate*						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	4.06	0.51	4.06	0.46	4.17	0.43
Dialogic	4.11	0.55	4.40	0.47	4.28	0.49
Fluency: *Number of False Starts per 100 Words*						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	0.72	0.80	0.30	0.48	0.41	0.61
Dialogic	0.92	1.02	1.07	0.96	1.19	1.10
Text Length (number of words)						
	Time one		Time two		Time three	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Monologic	152	68	156	62.65	148	64.15
Dialogic	179	89	174	101.96	172	79.3

Table 4.10 lists the correlations between the each of the above CAF measures and the oral proficiency scores and their respective *p*-values.

Table 4.10 Correlations between the CAF Measures and the Oral Proficiency Scores

Construct	CAF Measures	Correlations with the Oral Proficiency Scores	*p*-value
Syntactic complexity	Subordination Measure	0.231	<0.001*
	Mean Length of AS-unit	0.287	<0.001*
	Mean Length of Clause	0.238	<0.001*
	Coordination Measure	0.149	0.004*
	Number of Wh-clauses per 100 words	0.032	0.532
Accuracy	Number of error-free AS-units per 100 words	0.019	0.706
Fluency	Number of False Starts per100Words	-0.208	<0.001*
	Mean Length of Pauses	-0.344	<0.001*
	Mean Length Fluent Runs	0.563	<0.001*
	Articulation Rate	0.396	<0.001*
	Phonation-Time Ratio	0.448	<0.001*
	Text-length	0.422	<0.001*
Lexical Sophistication	MRC Familiarity All words	-0.142	0.006*
	MRC Meaningfulness All words	-0.178	<0.001*
	COCA Spoken Frequency All Words	-0.032	0.538

* $p < 0.05$

As Table 4.10 shows, in total, 11 CAF variables (four syntactic complexity variables [*Mean Length of AS-unit*, *Subordination Measure*, *Coordination Measure*, *Mean Length of Clause*], two lexical sophistication indices [*MRC Familiarity All Words*, *MRC Meaningfulness All Words*], and five fluency indices [*Mean Length of Pauses, Mean Length of Fluent Runs, Phonation-Time Ratio, Articulation Rate, Number of False Starts per 100 Words*]) had significant (p<0.05) correlations above 0.10 (r > 0.10) with the oral proficiency scores. The remaining three CAF variables (*Number of WH-Clauses per 100 Words, Number of Error-Free AS-units, and COCA Spoken Frequency All Words*) with correlations (with the response variable) below 0.10 (r <10) were discarded.

All the linguistic predictors were further analyzed for multicollinearity using their VIF scores, which measures the strength of linear relationship among independent variables. Higher VIF values result in inflated *p*-values leading to difficulty in interpreting results (Levshina, 2015). One syntactic complexity measure (*Mean Length of AS*-unit) and one fluency measure (*Mean Length of Pauses*) had VIF values higher than 4. After checking the correlations, *Mean Length of AS-unit* was found to be highly correlated with another syntactic complexity index, *Subordination Measure* (*r* = 0.90). Likewise, *Mean Length of Pauses* was found to have high correlation (*r* = 0.73) with another fluency measure, *Phonation-Time ratio*. Among these indices, *Mean Length of AS-unit* and *Phonation-Time Ratio* had higher correlation with the oral proficiency scores. Hence, *Subordination Measure* and *Mean Length of Pauses* were discarded from the analysis. None of the other linguistic predictors had VIF value greater than 2.

Additionally, after the goodness-of-fit tests with ANOVA, two syntactic complexity variables (*Mean Length of AS-unit*, *Coordination Measure*) and one lexical complexity index (*MRC Meaningfulness All Words*) were discarded. The remaining six CAF measures (*Mean Length of Clause, MRC Familiarity All Words, Mean Length of Fluent Runs, Phonation-Time Ratio, Articulation Rate, Number of False Starts per 100 Words*) along with text length, L1 distance, and program level (IEP/matriculated) were included as fixed factors in the LME models 1 (for monologic tasks) and 2 (for dialogic tasks).

Research question 1a focused on whether the relationships between CAF measures and L2 oral proficiency scores varied depending on task type (monologic/dialogic). To address this research question, two separate models were created: model 1 for monologic tasks and model 2 for dialogic tasks.

4.3.1 CAF predictors of oral proficiency for monologic tasks

Table 4.11 reports the LME model 1 for monologic tasks.

Table 4.11 Results of the LME model 1on the relationships between CAF based predictors of the monologic tasks and oral proficiency scores

					Random Effects			
	Fixed Effects				By participant		By prompt	
	Coef	Std. Error	*t*-value	*p*-value	Var	SD	Var	SD
Intercept	10.799	0.251	42.921	<0.001*	1.156	1.075	0.006	0.077
Phonation-time ratio	0.375	0.106	3.535	<0.001*	-	-	-	-
MRC Familiarity all words	-0.197	0.068	-2.886	0.004*	-	-	-	-
Mean length of clause	0.142	0.072	1.961	0.050*	-	-	-	-
L1 distance	0.351	0.151	2.314	0.024*	-	-	-	-
Program level: Matriculated	1.293	0.316	4.086	<0.001*	-	-	-	-
Text-length	0.405	0.078	5.149	<0.001*	-	-	-	-
False starts per 100 words	-0.200	0.069	-2.884	0.004*	-	-	-	-

Note. * $p < 0.05$
Var= Variance; SD= Standard Deviation; Coef =Coefficient

The findings in Table 4.11 show that in monologic tasks, *Phonation-Time Ratio* is significant positive predictor of L2 oral proficiency scores. This finding suggests that for each increase in *Phonation-Time Ratio* in monologic tasks, the participants' oral proficiency scores increased by 0.375. *Mean Length of Clause* is also significant with a positive coefficient, indicating that for each increase mean clause-lengths in monologic tasks, the participants' oral proficiency scores increased by 0.142. On the contrary, *False Starts per 100 Words* and *MRC Familiarity All Words* are significant negative predictors of the oral proficiency scores, indicating that for each increase in false starts and familiar vocabulary in the participants' monologic speech, their oral proficiency scores decreased by -0.200 and -0.197, respectively. The results also showed that *Program level: Matriculated* is a significant positive predictor of the oral proficiency scores indicating that compared to the IEP learners, the matriculated ESL learners' oral proficiency scores were higher by 1.293. Additionally, *Text-Length* was a significant predictor with a positive coefficient indicating that with every increase in the participants' speech length, their oral proficiency scores increased by 0.405.This model on the monologic tasks explains 45% variance in the oral proficiency scores (marginal R^2= 0.458, conditional R^2= 0.836).

4.3.2 CAF predictors of oral proficiency for dialogic tasks

Using similar analytical procedure (followed for creating the LME model 1 on monologic data), the LME model 2 was created for the dialogic task performance data. The output of the LME model 2 is reported in Table 4.12.

Table 4.12 Results of the LME model on the relationships between CAF based predictors of the dialogic tasks and oral proficiency scores

					Random Effects					
	Fixed Effects				By participant		By prompt		By pair-Combination	
	Coef	Std. Error	*t*-value	*p*-value	Var	SD	Var	SD	Var	SD
Intercept	10.902	0.283	38.43	<0.001*	1.49	1.22	<0.001	<0.001	0.07	0.27
False starts per 100 words	-0.17	0.07	-2.46	0.015*	-	-	-	-	-	-
Phonation-time ratio	0.214	0.085	2.516	0.012*	-	-	-	-	-	-
Articulation rate	0.212	0.099	2.142	0.033*	-	-	-	-	-	-
Text-length	0.452	0.091	4.953	<0.001*	-	-	-	-	-	-
Program level: Matriculated	1.110	0.359	3.090	0.003*	-	-	-	-	-	-

Note. * $p < 0.05$

Var= Variance; SD= Standard Deviation; Coef =Coefficient

As is reported in Table 4.12, *Phonation-Time Ratio* and *Articulation Rate* of dialogic task performance are significant positive predictors of oral proficiency scores, indicating that for each increase in *Phonation-Time Ratio* and *Articulation Rate* in the dialogic oral production, the participants' oral proficiency scores increased by 0.214 and 0.212, respectively. In contrast, *False Starts per 100 Words* was significant negative predictor, indicating that for each increase in false starts in the dialogic oral production, the participants' oral proficiency scores decreased by -0.17. The results also showed that *Program level: Matriculated* is a significant positive predictor of the oral proficiency scores indicating that compared to the IEP learners, the matriculated ESL learners' oral proficiency scores were higher by 1.110. Additionally, *Text-Length* was a significant predictor with a positive coefficient indicating that with every increase in the participants' speech length, their oral proficiency scores increased by 0.452. This model on the dialogic tasks explains 32% variance in the oral proficiency scores (marginal R^2= 0.325, conditional R^2= 0.835).

4.4 Research question 1b: Whether the relationships between the CAF measures and the oral proficiency scores varied depending on time (one/two/three)

To address the research question 1b (whether the relationships between the CAF measures of monologic and dialogic tasks and L2 oral proficiency scores varied depending on time [one/two/three]), main effects were checked, and interactions were fitted in model 1 (for monologic tasks) and model 2 (for dialogic tasks) between the CAF measures and time.

4.4.1 Interactions between time and the CAF predictors for the monologic tasks

Table 4.13 reports the model with significant interactions between the CAF measures of monologic tasks and time (with time: one as the reference). Time had a significant interaction with *Phonation-Time Ratio* in the monologic tasks. The negative coefficient (-0.285) of the significant predictor, "*Phonation-Time Ratio* x Time: three" indicates that at time three (compared to time one), for each increase in *Phonation-Time Ratio* in the participants' monologic speech, their oral proficiency scores decreased by -0.285. Thus, lower *Phonation-Time Ratio* in the monologic tasks was predictive of higher oral proficiency scores at time three compared to time one. This interaction model explained 45% variance in the oral proficiency scores (marginal R^2= 0.452, conditional R^2=0.844).

Table 4.13 Results of the LME model with Interactions between the CAF Measures of Monologic Tasks and Time

					Random Effects			
	Fixed Effects				By participant		By prompt	
	Coef	Std. Error	*t*-value	*p*-value	Var	SD	Var	SD
Intercept	11.091	0.270	41.066	<0.001*	1.23	1.11	<0.001	<0.001
Phonation-time ratio	0.673	0.122	5.495	<0.001*	-	-	-	-
Time three	-0.386	0.131	-2.934	0.004*	-	-	-	-
Time two	-0.400	0.130	-3.080	0.002*	-	-	-	-
MRC Familiarity all words	-0.236	0.066	-3.531	<0.001*	-	-	-	-
Text-length	0.371	0.077	4.795	<0.001*	-	-	-	-
L1 distance	0.340	0.156	2.182	0.033*	-	-	-	-
Program level: Matriculated	1.262	0.325	3.879	<0.001*	-	-	-	-
Phonation-time ratio x Time: three	-0.285	0.144	-1.975	0.048*	-	-	-	-
Phonation-time ratio x Time: two	-0.154	0.126	-1.220	0.222	-	-	-	-
Phonation-time ratio x Time: two versus three	-0.131	0.141	0.925	0.354	-	-	-	-

Note. * $p < 0.05$

Var= Variance; SD= Standard Deviation; Coef =Coefficient
Time: one = Reference level for the 'Time' variable

4.4.2 Interactions between time and the CAF predictors of the dialogic tasks

Table 4.14 reports the output of the model that fitted interactions between the CAF measures of the dialogic tasks and time. The results showed significant negative interactions between *Phonation-Time Ratio* of dialogic oral production and time and between *Articulation Rate* of dialogic oral production and time. Additionally, the interaction model reports significant positive interactions between *Mean Length of Fluent Runs* of dialogic oral production and time. This interaction model explained about 42% variance in the oral proficiency scores (marginal R^2= 0.427, conditional R^2= 0.845).

Table 4.14 Results of the LME model with Significant Interactions between the CAF Measures of Dialogic Tasks and Time

					Random Effects					
	Fixed Effects				By participant		By prompt		By pair-Combination	
	Coef	Std. Error	*t*-value	*p*-value	Var	SD	Var	SD	Var	SD
Intercept	11.21	0.286	39.07	<0.001*	1.347	1.16	<0.001	<0.001	<0.001	<0.001
Mean length of clause	0.154	0.074	2.088	0.038*	-	-	-	-	-	-
Phonation-time ratio	0.608	0.131	4.626	<0.001*	-	-	-	-	-	-
Time: three	-0.379	0.143	-2.65	0.009*	-	-	-	-	-	-
Time: two	-0.419	0.146	-2.86	0.005*	-	-	-	-	-	-
Mean length of fluent runs	-0.322	0.183	-1.76	0.081	-	-	-	-	-	-
Articulation rate	0.494	0.129	3.833	<0.001*	-	-	-	-	-	-
Text-length	0.321	0.095	3.384	<0.001*	-	-	-	-	-	-
L1 distance	0.310	0.165	1.975	0.050*	-	-	-	-	-	-
Program level: Matriculated	1.047	0.344	3.045	0.003*	-	-	-	-	-	-
Phonation-time ratio x time: three	-0.691	0.226	-3.05	0.002*	-	-	-	-	-	-
Phonation-time ratio x time: two	-0.429	0.194	-2.21	0.028*	-	-	-	-	-	-

Phonation-time ratio x time: two versus three	-0.262	0.234	-1.12	0.262	-	-	-	-	-	-
Mean length of fluent runs x time: three	0.473	0.228	2.075	0.040*	-	-	-	-	-	-
Mean length of fluent runs x time: two	0.574	0.225	2.553	0.012*	-	-	-	-	-	-
Mean length of fluent runs x time: two versus three	-0.101	0.211	-0.48	0.632	-	-	-	-	-	-
Articulation rate x Time: three	-0.342	0.157	-2.18	0.031*	-	-	-	-	-	-
Articulation rate x Time: two	-0.125	0.163	-0.77	0.442	-	-	-	-	-	-
Articulation rate x Time: two versus three	-0.216	0.167	-1.29	0.197	-	-	-	-	-	-

Note. * $p < 0.05$
Var= Variance; SD= Standard Deviation; Coef =Coefficient
Time: one = Reference level for the 'Time' variable

As can be seen in Table 4.14, time had significant negative interactions with the fluency measure, *Phonation-Time Ratio* in the dialogic tasks. The negative coefficient (-0.691) of the significant predictor, "*Time three* x *Phonation-Time Ratio*" indicates that at time three (compared to time one), for each increase in *Phonation-Time Ratio* in dialogic task performance, the participants' oral proficiency scores decreased by -0.691. Similarly, the negative coefficient (-0.429) of the significant predictor "*Phonation-Time Ratio x Time: two*" indicates that at time two (compared to time one), for each increase in *Phonation-Time Ratio* in dialogic oral production, the participants' oral proficiency scores decreased by -0.429. Figure 4.1 visually displays these interaction effects.

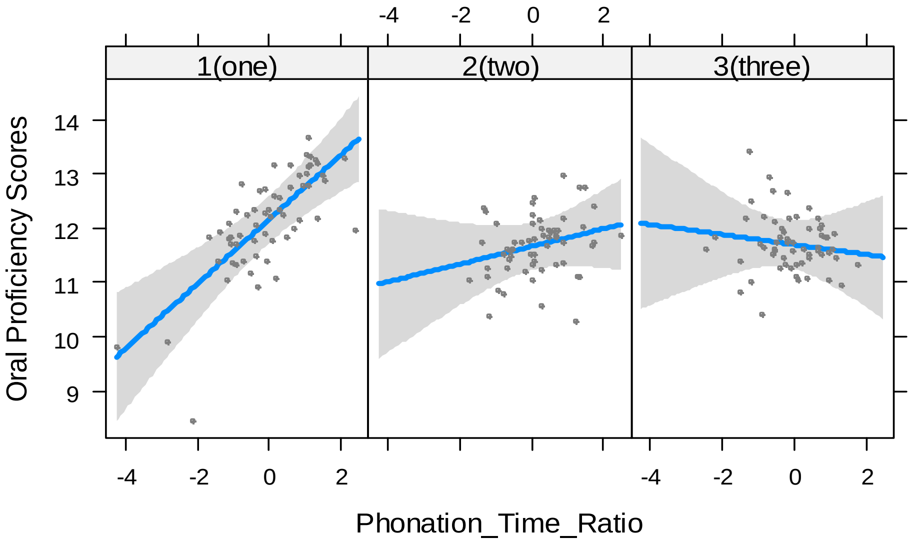

Figure 4.1 The Plot of Significant Interactions between "Phonation-Time Ratio" in Dialogic Speech and Time in their Effects on the Oral Proficiency Scores

As Figure 4.1 shows, at time one, for higher *Phonation-Time Ratio* in dialogic speech, the participants' oral proficiency scores increased. However, compared to time one, at time two and time three, for higher *Phonation-Time Ratio* in dialogic task performances, the participants had lower oral proficiency scores. Hence, Lower *Phonation-Time Ratio* (reduced phonation in proportion to their total duration of speech) in dialogic oral production was predictive of higher L2 oral proficiency scores at time two and time three compared to time one.

Similarly, as Table 4.14 shows, the negative coefficient (-0.342) of the significant predictor "*Articulation rate* x Time: three" indicates that at time three (compared to time one), for each increase in *Articulation Rate* in the participants' dialogic speech, their oral proficiency scores decreased by -0.342. Thus, lower *Articulation Rate* in dialogic oral production was predictive of higher oral proficiency longitudinally (over eight months). This interaction between time (with time: one as the reference level) and *Articulation Rate* is visually displayed in Figure 4.2.

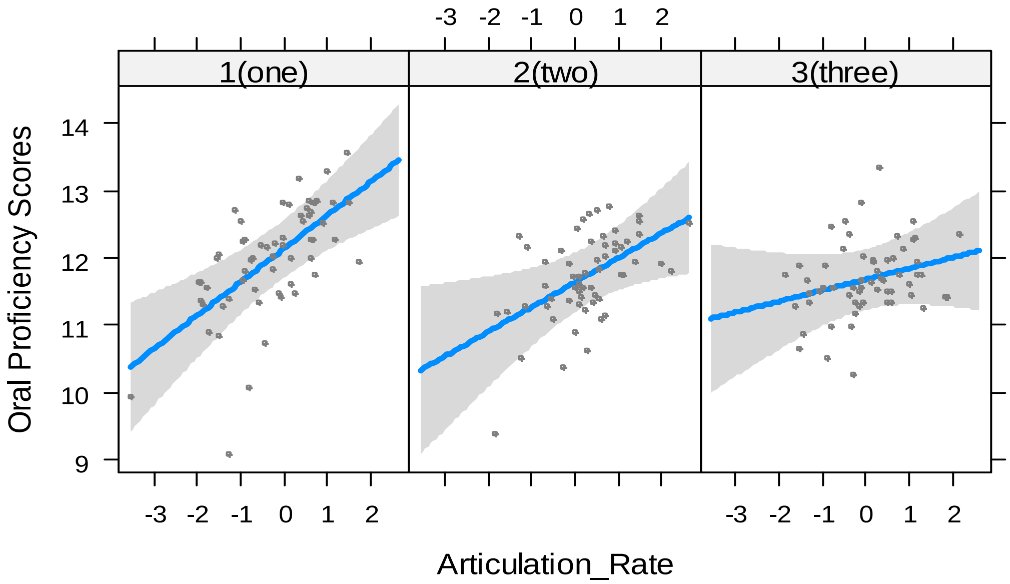

Figure 4.2 The Plot of Significant Interactions between "Articulation Rate" in dialogic speech and Time in their Effects on the Oral Proficiency Scores

As is shown in Figure 4.2, at time three, compared to time one, the participants with higher *Articulation Rate* in dialogic speech had lower oral proficiency scores. Thus, lower *Articulation Rate* in dialogic speech was predictive of higher oral proficiency scores at time three compared to time one.

In contrast, as is also reported in Table 4.14, the breakdown fluency measure, *Mean Length of Fluent Runs*, had significant positive interactions with time in their effects on the oral proficiency scores. The positive coefficients of the significant predictors, "*Mean Length of Fluent Runs* x Time: two" and "*Mean Length of Fluent Runs* x Time: three" indicate that at time two and time three (compared to time one), for each increase in *Mean Length of Fluent Runs* in dialogic oral production, the participants' oral proficiency scores also increased by 0.574 and 0.473, respectively. Thus, longer fluent runs in dialogic speech was predictive of higher L2 oral proficiency scores at time two and time three compared to time one. A visual display of these interaction effects in presented in Figure 4.3.

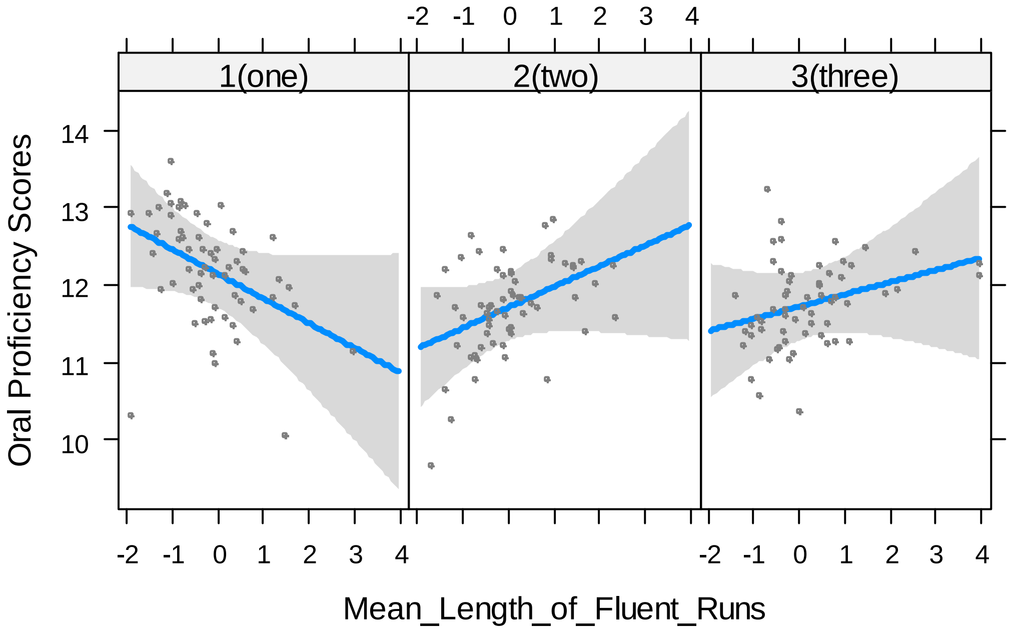

Figure 4.3 The Plot of Significant Interactions between "Mean Length of Fluent Runs" in Dialogic Speech and Time in their Effects on the Oral Proficiency Scores

As can be seen in Figure 4.3, at time two and time three, the participants with higher oral proficiency scores produced significantly longer fluent runs compared to time one. Hence, longer fluent runs in dialogic speech was predictive of higher L2 oral proficiency scores longitudinally over eight months.

4.5 Research question 2a: ESL speakers' WM and aptitude measures as predictors of their oral proficiency scores

To answer the research question 2a (whether ESL speakers' WM and aptitude measures predict their oral proficiency scores), *Explicit Aptitude* (average of LLAMA E and F), SRT scores, LLAMA B, and LLAMA D were included as the aptitude measures. As the *EWM* measure, summation of the operation span and the symmetry span scores and as the PM measure, the digit span scores were used. Table 4.15 presents the correlations between the EWM and aptitude variables, and Table 4.16 reports the correlations between the PM and the aptitude variables.

Table 4.15 Correlations between the EWM and the Aptitude Measures

Variables	Correlations with the EWM scores	*p* value
Explicit Aptitude (Average of LLAMA E and LLAMA F)	0.378	0.002*
SRT	0.087	0.507
LLAMA B	0.313	0.014*
LLAMA D	0.199	0.126

**p* < 0.05

Table 4.16 Correlations between the PM and the Aptitude Measures

Variables	Correlations with the PM scores	*p* value
Explicit Aptitude (Average of LLAMA E and LLAMA F)	0.344	0.007*
SRT	0.117	0.371
LLAMA B	0.282	0.028*
LLAMA D	0.291	0.024*

**p* < 0.05

Table 4.17 presents descriptive statistics of the WM and aptitude measures, and Table 4.18 presents the correlations between the WM and aptitude measures and the oral proficiency scores.

Table 4.17 Descriptive Statistics of the Aptitude and WM Measures

	Total possible score	Mean	Std. Dev	Min	Max
Explicit Aptitude (Average of LLAMA E and LLAMA F)	100	64	20.15	5	95
SRT score	6	1.78	1.37	0	6
EWM (Summation of Operation and Symmetry Span scores)	39	26.97	6.70	9	39
PM (digit span test) score	168	130.08	34.03	60	168
LLAMA B	100	47.33	19.05	15	85

LLAMA D	100	25.42	12.99	0	60

Note. Std. Dev= Standard deviation; Min= Minimum; Max= Maximum

Table 4.18 Correlations between the WM and Aptitude Variables and the Oral Proficiency Scores

WM/Aptitude Variables	Correlation with the Oral Proficiency Scores	*p*-value
EWM scores (summation of operation and symmetry span scores)	0.110	0.129
Explicit Aptitude (average of LLAMA E and LLAMA F)	0.212	0.004*
SRT score	-0.046	0.535
PM (digit span test) score[9]	0.037	0.616
LLAMA B	0.030	0.681
LLAMA D	0.026	0.726

**p <0.05

As can be seen in Table 4.18, SRT, PM, EWM, LLAMA B, and LLAMA D scores had non-significant correlations with the oral proficiency scores (dependent variable), and hence, these variables were discarded. Only *Explicit Aptitude* has a statistically significant correlation with the oral proficiency scores. Hence, *Explicit Aptitude* (average of LLAMA E and LLAMA F scores) was selected to enter the LME model 3 as a predictor along with L1 distance, and program level (IEP/matriculated). Table 4.19 reports the output of the LME model 3. This main model explained 23% variance in the oral proficiency scores (marginal R^2= 0.232, conditional R^2=0.824). As can be seen in Table 4.19, *Explicit Aptitude* was not a significant predictor of the oral proficiency scores. Thus, neither aptitude nor WM predicted L2 oral proficiency.

[9] To examine whether participants of the same first language have stronger correlation between their digit span scores and oral proficiency, two additional correlations were run: one correlation was between the digit span scores of only the Chinese speakers (n=23) and their oral proficiency scores and another was between the digit span scores of only the Arabic speakers (n=13) and their oral proficiency scores. However, neither analysis showed significant correlations (r = 0.09, p>0.05 for Arabic; r = -0.03, p>0.05 for Chinese).

Table 4.19 Output of the LME Model 3 with Explicit Aptitude as the Predictor

					Random Effects	
	Fixed Effects				By participant	
	Coefficient	Std. Error	*t*-value	*p*-value	Variance	SD
Intercept	10.741	0.368	29.151	<0.001*	2.201	1.483
Explicit aptitude	0.273	0.235	1.164	0.249	-	-
L1 Distance	0.464	0.207	2.242	0.028*	-	-
Program Level: Matriculated	1.382	0.487	2.834	0.006*	-	-

Note. IEP= Reference for the "program level" variable
SD= Standard Deviation
* p < 0.05

4.6 Research question 2b: Mediating effects of time on the relationships between the WM and aptitude measures and the oral proficiency scores

To answer the question on whether the relationships between the WM and aptitude variables and the oral proficiency scores change over time, an interaction model was fitted with interactions between the *Explicit Aptitude* (because only *Explicit Aptitude* had significant correlation with the dependent variable [oral proficiency scores]) and time (one/two/three). Table 4.20 reports the output of the interaction model. This model explained 23% variance in the oral proficiency scores (marginal R^2=0.235, conditional R^2= 0.824). As can be seen in Table 4.20, time had no significant interactions with the *Explicit Aptitude* scores in their effects on L2 oral proficiency.

Table 4.20 Output of the LME model with Interactions between Explicit Aptitude and Time

					Random Effects	
	Fixed Effects				By participant	
	Coefficient	Std. Error	*t*-value	*p*-value	Variance	SD
Intercept	10.887	0.378	28.785	<0.001*	2.201	1.483
Time three	-0.224	0.147	-1.524	0.130	-	-
Time two	-0.210	0.147	-1.428	0.155	-	-

Explicit Aptitude	0.354	0.250	1.416	0.161	-	-
L1 distance	0.464	0.207	2.242	0.028*	-	-
Program level: matriculated	1.382	0.487	2.834	0.006*	-	-
Time three x Explicit Aptitude	-0.149	0.148	-1.011	0.313	-	-
Time two x Explicit Aptitude	-0.091	0.148	-0.620	0.536	-	-
Time two vs. three x Explicit Aptitude	-0.057	0.148	0.392	0.695	-	-

Note. IEP= Reference for the “program level” variable
Time one= Reference for the “time” variable
SD= Standard Deviation
* $p < 0.05$

4.7 Research question 3: Whether the relationships between the CAF measures and L2 oral proficiency scores are mediated by ESL speakers’ WM and aptitude

4.7.1 Interactions between Explicit Aptitude and the CAF predictors of the monologic tasks

To answer the research question 3 (whether the relationships between the CAF measures and the oral proficiency scores are mediated by the ESL speakers’ WM and aptitude abilities), the LME model 4 was developed with interactions between the CAF measures of the monologic tasks (those selected for answering the questions 1a and 1b) and the individual difference variable included in the models for answering the questions 2a and 2b (*Explicit Aptitude* because only *Explicit Aptitude* had significant correlations with the oral proficiency scores). This model showed significant interactions between *MRC Familiarity All Words* and *Explicit Aptitude.* Table 4.21 reports the output of the LME model 4 with the significant interaction. This model explained 50% variance in the oral proficiency scores (marginal R^2= 0.501, conditional R^2= 0.840).

Table 4.21 Output of the LME Model with Significant Interactions between the CAF Measures of the Monologic Tasks and Explicit Aptitude

					Random Effects			
	Fixed Effects				By participant		By prompt	
	Coef	Std. error	*t*-value	*p*-value	Var	SD	Var	SD
Intercept	10.969	0.264	41.407	<0.001*	1.051	1.025	<0.001	<0.001
Mean length of clause	0.146	0.071	2.031	0.044*	-	-	-	-
Phonation-time ratio	0.399	0.103	3.867	<0.001*	-	-	-	-
Text-length	0.363	0.077	4.717	<0.001*	-	-	-	-
MRC Familiarity all words	-0.192	0.066	-2.89	0.004*	-	-	-	-
Explicit Aptitude	0.199	0.169	1.178	0.244	-	-	-	-
L1 distance	0.349	0.149	2.344	0.022*	-	-	-	-
Program level: Matriculated	1.086	0.352	3.078	0.003*	-	-	-	-
MRC Familiarity all words x Explicit Aptitude	0.250	0.064	3.860	<0.001*	-	-	-	-

Note. * $p < 0.05$
Var= Variance; SD= Standard Deviation; Coef =Coefficient

In Table 4.21, the positive coefficient (0.250) of the significant predictor, *MRC Familiarity All Words* x *Explicit Aptitude,* indicates that for each increase in *Explicit Aptitude* and familiar vocabulary in monologic speech, the participants' oral proficiency scores increased by 0.250. Thus, the participants with higher *Explicit Aptitude* and higher oral proficiency used more familiar vocabulary in their monologic task performances. The significant interactions between *Explicit Aptitude* and the lexical sophistication measure, *MRC Familiarity All Words* is visually displayed in Figure 4.4. Since *Explicit Aptitude* is a continuous variable, in all the interaction plots with *Explicit Aptitude* as a predictor, three upper panels are set up for aptitude using the 10th, 50th, and 90th percentiles (corresponding to the standardized aptitude scores -1.4, 0.3, and 1.3, respectively, in Figure 4.4) (Breheny, 2020). In all the Figures, "Aptitude" refers to *Explicit Aptitude*.

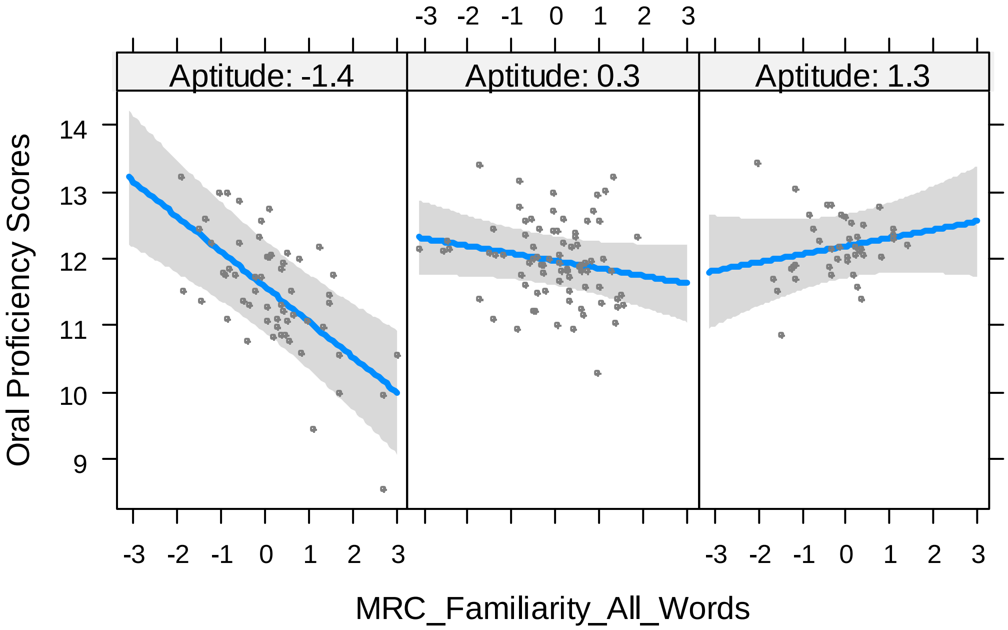

Figure 4.4 The Plot of Significant Interactions between "MRC Familiarity All Words" in the Monologic Tasks and Explicit Aptitude in their Effects on the Oral Proficiency Scores

As shown in Figure 4.4, for the participants with higher *Explicit Aptitude* (e.g., above the 90th percentile in the rightmost panel), for each increase in word familiarity scores, their oral proficiency scores also increased. Thus, the participants with higher *Explicit Aptitude* and higher oral proficiency used more familiar vocabulary in monologic oral production.

4.7.2 Interactions between Explicit Aptitude and the CAF predictors of the dialogic tasks

Another LME model (model 5) was developed with interactions between the CAF measures of the dialogic tasks (those selected for answering the questions 1a and 1b) and *Explicit Aptitude* to answer the research question 3 (whether the relationships between the CAF measures and the oral proficiency scores are mediated by the ESL speakers' WM and aptitude abilities). This model reported significant negative interactions between *Mean Length of Clause* of dialogic oral production and *Explicit aptitude* and between *Mean Length of Fluent Runs* of dialogic oral production and *Explicit Aptitude*. This interaction model explains 46% variance in the oral proficiency scores (marginal R^2= 0.461, conditional R^2= 0.850). Table 4.22 reports this interaction model for the dialogic tasks with the significant interactions.

Table 4.22 Output of the LME Model 4 with Significant Interactions between the CAF Measures of the Dialogic Tasks and Explicit Aptitude

					Random Effects					
	Fixed Effects				*By participant*		*By prompt*		*By pair-combination*	
	Coef	Std. error	*t*-value	*p*-value	*Var*	*SD*	*Var*	*SD*	*Var*	*SD*
Intercept	11.194	0.282	39.649	<0.001*	1.249	1.117	<0.001	0.003	<0.001	<0.001
Mean length of clause	0.054	0.074	0.734	0.464	-	-	-	-	-	-
Explicit Aptitude	0.205	0.179	1.147	0.256	-	-	-	-	-	-
Phonation-time ratio	0.206	0.088	2.328	0.021*	-	-	-	-	-	-
False starts per 100 words	-0.199	0.064	-3.10	0.002*	-	-	-	-	-	-
Mean length of fluent runs	-0.005	0.110	-0.05	0.960	-	-	-	-	-	-
Articulation rate	0.209	0.096	2.172	0.031*	-	-	-	-	-	-
Text-length	0.420	0.089	4.700	<0.001*	-	-	-	-	-	-

Program level: Matriculated	0.790	0.373	2.119	0.038 *	-	-	-	-	-	-
Mean length of clause x Explicit Aptitude	-0.21	0.073	-2.85	0.004 *	-	-	-	-	-	-
Mean length of fluent runs x Explicit Aptitude	-0.299	0.103	-2.90	0.004 *	-	-	-	-	-	-

Note. * $p < 0.05$
Var= Variance; SD= Standard Deviation; Coef =Coefficient

As Table 4.22 shows, the negative coefficient (-0.299) of the significant predictor, *Mean Length of Fluent Runs x Explicit Aptitude* indicates that for every increase in the participants' *Explicit Aptitude* and *Mean Length of Fluent runs* in their dialogic speech, their oral proficiency scores decreased by -0.299 (see Figure 4.5 for a visual display of this interaction effect). This finding suggests that high proficiency participants with higher *Explicit Aptitude* had lower *Mean Length of Fluent Runs* in dialogic speech.

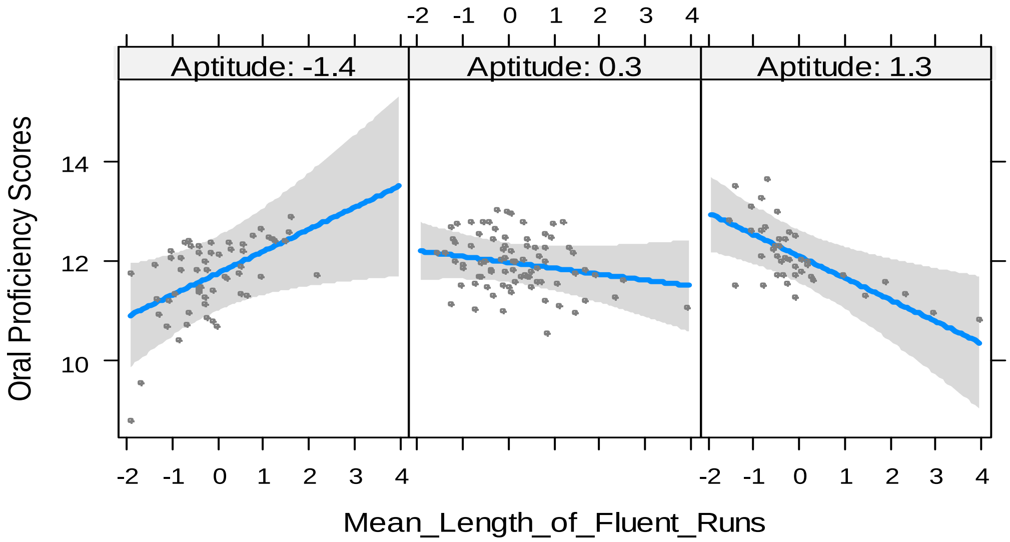

Figure 4.5 The Plot of Significant Interactions between "Mean Length of Fluent Runs" in the Dialogic Tasks and Explicit Aptitude in their Effects on the Oral Proficiency Scores

Figure 4.5 shows that in the dialogic speech, for the participants with higher *Explicit Aptitude* (as shown in the two right panels), for higher length of fluent runs, their oral proficiency scores decreased. Therefore, high proficiency ESL speakers with higher *Explicit Aptitude* produced shorter fluent runs in the dialogic tasks.

Likewise, the negative coefficient (-0.21) of the significant predictor, *Mean Length of Clause* x *Explicit Aptitude* suggests that for every increase in the participants' *Explicit Aptitude* and *Mean Length of Clause* in dialogic tasks, their oral proficiency scores significantly decreased by -0.210 (see Figure 4.6 for a visual display of this interaction effect). Thus, high proficiency participants with higher *Explicit Aptitude* had shorter *Mean Length of Clause* in the dialogic tasks.

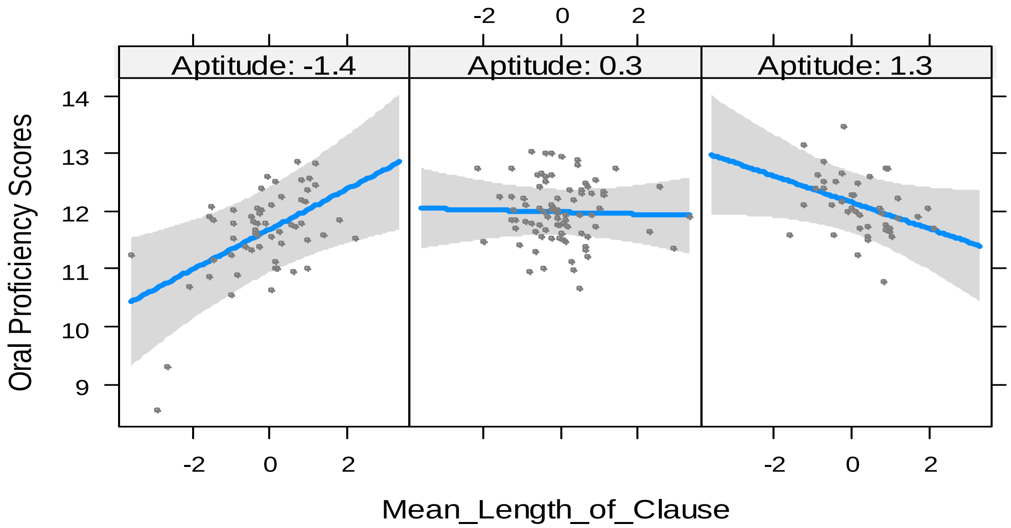

Figure 4.6 The Plot of Significant Interactions between Mean Length of Clause in the Dialogic Tasks and Explicit Aptitude in their Effects on the Oral Proficiency Scores

As shown in Figure 4.6, in the dialogic tasks, for the participants with higher aptitude (e.g., at 90th percentile, the rightmost panel), for increase in mean clause-lengths, their oral proficiency scores decreased. This finding suggests that the participants with higher oral proficiency and higher *Explicit Aptitude* used significantly shorter clauses in the dialogic tasks.

4.8 Summary of the results

The purpose of the dissertation was to investigate how the relationships between CAF measures of oral production and L2 oral proficiency vary depending on task-type (e.g., monologic versus dialogic) and time (one/two/three). The study also examined whether ESL speakers' individual differences in WM and aptitude were predictive of variations in their oral proficiency over time. Additionally, the dissertation investigated the interactions between ESL speakers' WM/aptitude and CAF measures of their oral productions in their combined effects on L2 oral proficiency. The results are summarized in Table 4.23.

Table 4.23 Summary of the Results of the Dissertation

Research Question 1: The effects of task type (monologic/dialogic) and time (one/two/three) on the relationships between CAF measures and L2 oral proficiency		
Sub-questions	Significant results found?	Significant predictors
1(a) Do the relationships between the CAF measures and L2 oral proficiency vary depending on task type (monologic/dialogic)?	yes	**Similarities:** In both monologic and dialogic tasks, higher breakdown fluency (*Phonation-Time Ratio*) and higher repair fluency (lower *False Starts per 100 Words*) were predictive of higher oral proficiency. **Differences:** In monologic speech, higher phrasal complexity (*Mean Length of Clause*) and higher lexical sophistication (lower scores in *MRC Familiarity All Words*), whereas in dialogic speech, higher speed fluency (*Articulation Rate*) were predictive of higher oral proficiency.
1(b) Do the relationships between the CAF measures and L2 oral proficiency vary depending on time (one/two/three)?	yes	**Monologic task:** The participants with higher oral proficiency had lower *Phonation-Time Ratio* at time three compared to time one. **Dialogic task**: The participants with higher oral proficiency had higher *Mean Length of Fluent Runs* at time three and time two compared to time one. In

		contrast, the participants with higher oral proficiency had lower *Phonation-Time Ratio* at time three and time two compared to time one. Likewise, lower *Articulation Rate* was predictive of higher oral proficiency scores at time three (compared to time one).
Research question 2: Relationships between WM and aptitude and L2 oral proficiency		
Sub-questions	Significant results found?	Significant predictors
2(a) Do WM and Aptitude predict L2 oral proficiency?	No	None
2(b) Do WM/aptitude predict variations in L2 oral proficiency over time?	No	None
Research question 3: Mediating effects of WM/aptitude on the relationships between CAF measures of oral production and L2 oral proficiency		
Sub-questions	Significant results found?	Significant predictors
3. Is there any mediating effects of WM/aptitude on the relationships between the CAF measures of monologic and dialogic tasks and L2 oral proficiency?	Yes	**Monologic tasks**: Participants with higher *Explicit Aptitude* and higher oral proficiency used more familiar vocabulary in monologic speech. **Dialogic tasks**: Participants with higher *Explicit Aptitude* and higher oral proficiency used shorter *Mean Length of Fluent Runs* and shorter *Mean Length of Clauses* in the dialogic speech.

Overall, the findings showed that in monologic speech, the ESL speakers with higher oral proficiency used longer clauses and less familiar vocabulary, whereas in the dialogic speech, the participants with higher oral proficiency had higher articulation rate (speed fluency). Additionally, regarding the mediating effects of time, the present findings show that in the monologic tasks, high proficiency ESL speakers produced lower *Phonation-Time Ratio* at time three compared to time one. Similarly, in dialogic speech, ESL speakers with higher oral proficiency also produced lower *Phonation-Time Ratio* at time two and time three compared to time one. Likewise, lower *Articulation Rate* in dialogic speech predicted higher oral proficiency scores at time three compared to time one. In contrast, high proficient ESL speakers produced longer *Mean Length of Fluent Runs* in dialogic speech at time two and time three compared to time one.

About the effects of individual cognitive differences on L2 oral proficiency, the current study did not find *Explicit Aptitude*, *Implicit Aptitude,* or any of the WM measures as significant predictors of the oral proficiency scores (although only *Explicit Aptitude* had significant correlations with the oral proficiency scores). *Explicit Aptitude* also did not predict any variation in L2 oral proficiency over time.

However, the participants' *Explicit Aptitude* had significant interactions with the CAF measures of monologic and dialogic oral production in their effects on L2 oral proficiency. The findings show that in monologic speech, the participants with higher *Explicit Aptitude* and higher oral proficiency used more familiar vocabulary. Additionally, in the dialogic speech, the participants with higher *Explicit Aptitude* and higher oral proficiency used shorter fluent runs and shorter clauses.

5 DISCUSSION

5.1 Research question 1a: Whether the relationships between the CAF measures and L2 oral proficiency scores vary depending on task type (monologic versus dialogic)

The findings of the question 1a (whether the relationships between the CAF measures and L2 oral proficiency scores vary depending on task type) suggest both similarities and differences in CAF based predictors of L2oral proficiency between monologic and dialogic tasks. Regarding the similarities, the participants with higher oral proficiency had higher phonation compared to pauses (higher *Phonation-Time Ratio*) and fewer false starts in both monologic and dialogic tasks. These findings are in line with Ferrari (2012) who also found that with development in proficiency over time, L2 learners of Italian produced speech with less pauses and higher phonation in both monologic and dialogic tasks. The participants with higher oral proficiency in the dissertation also produced fewer false starts in both monologic and dialogic speech. Speaking involves real-time decision making while transforming ideas from thought into speech (Segalowitz, 2000). The process of making such online decisions may lead to false starts during speech production (Segalowitz, 2000, p. 201). L2 speakers with higher oral proficiency may perform such real-time decision-making during speech production more efficiently. Hence, they have less false starts (in both monologic and dialogic speech) than those with lower proficiency.

In contrast, the results also showed differences in CAF-based predictors of oral proficiency between monologic and dialogic tasks. Firstly, the participants with higher oral proficiency produced longer clauses and unfamiliar vocabulary in monologic tasks. However, these two indices were not significant predictors of the oral proficiency scores for the dialogic tasks. Michel et al. (2007) and Michel (2011) also found that L2 learners of Dutch produced more complex language in monologic speech compared to that in dialogic speech. Similarly, Ferrari (2012) found that Italian as L2 learners produced longer clauses in monologic tasks than in dialogic tasks. As L2 speakers develop proficiency, they tend to use less subordination and more phrasal elaboration strategies (e.g., by modifiers) as a gradual move from dynamic (characterized by coordination and subordination) to synoptic (characterized by nominalizations and grammatical metaphors) complexification strategies (Ortega, 2012). Grammatical metaphors affect the lengths of elements inside clauses, and hence, it leads to longer phrases and longer individual clauses (Ortega, 2012). The present findings also show that in monologic speech, high proficient ESL speakers used longer clauses. Similarly, the findings also showed that ESL speakers with higher oral proficiency produced less familiar words in monologic speech. Vocabulary that are less familiar to ESL speakers might also be more difficult, which is indicative of proficient vocabulary usage (Salsbury et al., 2011). This finding could be compared to previous research on L2 writing (e.g., Crossley et al., 2012) although such a comparison needs to be considered with caution because of the difference in modality (speaking versus writing). Similar to the present finding, Crossley et al. (2012) found that as L2 learners' proficiency level increases, they use less familiar (and less imageable and more infrequent) words in L2 writing. Likewise, Crossley and Salsbury (2011) showed that ESL learners of higher proficiency used less familiar lexical bundles. However, *Mean Length of Clause* and *MRC Familiarity All Words* were not significant predictors of L2 oral proficiency in dialogic speech. These variations in CAF predictors of L2 oral proficiency between monologic and dialogic tasks might be explained with reference to the different information processing demands of monologic versus dialogic tasks.

There are different cognitive processes underlying monologic and dialogic task performances (Michel, 2011). Speech production in monologic tasks is dependent on the knowledge and cognitive resources of the speakers themselves (Michel, 2011). Monologic tasks are non-interactive with no listening involved (Robinson, 2001). In contrast, dialogic tasks involve interactions which entail listening to an interlocutor, that necessitates transformation of information from speech sounds into thoughts (Segalowitz, 2000) and formulating appropriate responses (De Jong & Perfetti, 2011). Also, in dialogic interactions, as is argued in the Interaction Approach, a participant’s speech can be impacted by the extent to which their interlocutors understood their speech; for example, speakers might need to address clarification requests and produce comprehension checks (Gass & Mackey, 2015). Moreover, as dialogic interactions are situated in social contexts, non-verbal cues or gestures of interlocutors may also impact L2 interactive speech (Sime, 2008); for example, in response to an interlocutor's gesture indicating incomprehension, a speaker may need to offer further explanations. Therefore, there are several variables that can affect an L2 speaker's regular flow of speech in a dialogic interaction. The process of speaking is complex in itself because of the online decision-making demands (Segalowitz, 2000). When such a complex process is accompanied with the demands of a dialogic task (e.g., processing interlocutor's arguments and producing appropriate arguments/counter-arguments in response), it might lead to production of shorter turns (Robinson, 2001) and less complex language.

Table 5.1 shows samples of monologic and dialogic speech (from time one) samples from a participant (Participant E) with relatively higher oral proficiency (oral proficiency score =13.4, which is above the average [11.76, as was reported in Table 4.5]). Table 5.1 displays the first 20 seconds of transcribed monologic and dialogic speech from participant E.

Table 5.1 Samples of Monologic and Dialogic Speech from a Relatively Higher Proficient Participant

Participant: E
Oral Proficiency score: 13.4/16

	• Monologic (version E)		• Dialogic (version A)
1	|*I would like ::to live in Atlanta downtown ::when I am attending Georgia State University ::because it is more convenient.*|		**Partner**:…
		1	**E**: |*Have you decided the smaller picture or just*|
			Partner:…
2	|*and if you are living away from university ::you will take so much time ::to come back ::that maybe you would not enjoy all of those opportunities ::you have*|		**E**: {*this is for an*}
		2	|*okay undergraduate studies,*|
		3	|*okay I will disagree with you and then go*|
			Partner:..
		4	**E**: |*that's on*|
			Partner:...
		5	**E**: |*small*|
			Partner:…
		6	**E**: |*why small living?*|
			Partner:…
		7	**E**: |*Well I prefer*|

Note. In this table, notations of AS-unit analysis (Foster et al., 2001) are used to indicate AS-units (enclosed in two upright slashes,| |), clauses (divided by two double colons, ::), and false starts (within curly brackets {}). For the dialogic speech, the partner's speech is not included. Every AS-unit is numbered.

As is shown in Table 5.1, in the monologic speech, the participant 'E' is producing clauses such as "*if you are living away from university"* (AS-unit 2)*, "because it is more convenient"* (AS-unit 1)*, "that maybe you would not enjoy all of those opportunities"* (AS-unit 2) that are longer (including efficient use of subordinating conjunctions, such as, "if", "because") than those produced in the dialogic speech. In the dialogic speech, there are multiple uses of noticeably shorter clauses, for example, "*that's on*" (AS-unit 4), "*why small living?*" (AS-unit 6), and "well, *I prefer"* (AS-unit 7). In the dialogic excerpt, the participant 'E' engages in frequent interactions with her partner, which might have influenced her use of shorter clauses.

Additionally, the findings showed that in dialogic tasks, in contrast to monologic tasks, high proficient ESL speakers had higher *Articulation Rate* (that measures speed fluency). This finding supports those of previous studies (e.g., Ferrari, 2012; Tavakoli, 2016; Michel et al., 2007, Michel, 2011) that also found higher fluency in L2 dialogic speech than in monologic speech. In Tavakoli (2016), ESL learners had significantly faster articulation rates in dialogues than in monologues. This finding underscores the argument that having a partner in dialogic speech might encourage high proficient L2 speakers to communicate interactively and address the interlocutors' needs by producing faster speech and fewer hesitations (Tavakoli, 2016). Moreover, during dialogic interactions, L2 speakers with higher oral proficiency can use their partner's turn to plan for their upcoming utterances (Webber, 2008). Thus, listening to the interlocutors might help high proficient ESL speakers to better conceptualize (generating preverbal message, Levelt, 1989) and reformulate (converting the preverbal message to a phonetic plan for speech, Levelt, 1989) their utterances and produce faster speech (Tavakoli, 2016).

5.2 Research question 1b: Mediating effects of time on the relationships between the CAF measures and the oral proficiency scores

The results of the research question 1b (Whether the relationships between the CAF measures of monologic and dialogic tasks and the oral proficiency scores change over time) show that higher *Mean Length of Fluent Runs* in dialogic speech was predictive of higher oral proficiency scores over time. Thus, *Mean Lengths of Fluent Runs* in dialogic speech was a significant predictor of development in L2 oral proficiency scores from time one to time two and time three (over eight months). *Mean Length of Fluent Runs* indicates average length of the speaking turns in-between pauses, and thus, over eight months, ESL speakers with higher oral proficiency produced significantly longer runs between pauses. This finding supports that of Tonkyn (2012) who examined changes in CAF measures of dialogic speech produced by upper intermediate instructed learners of English over 9-weeks. Tonkyn (2012) found that over this period, ESL learners produced significantly longer fluent runs in dialogic discussions. Likewise, in Tavakoli (2016), ESL learners had higher fluency in dialogic speech than in monologic speech because the participants significantly produced longer fluent runs in dialogues than in monologues. The interactive nature of dialogic tasks might encourage high proficiency ESL speakers' willingness to communicate (Tavakoli, 2016), which might have led to their use of longer fluent runs in dialogic speech.

In contrast, the participants with higher oral proficiency had significantly lower *Phonation-Time Ratio* in both monologic and dialogic speech at time three compared to time one. *Phonation-Time Ratio* is related to the number of pauses in speech, and "if the mean length of pauses is stable but the number of pauses decreases, phonation/time ratio increases" (De Jong & Perfetti, 2011, p. 538). Thus, lower *Phonation-Time Ratio* indicates higher frequency of pauses in speech. Therefore, high proficient L2 speakers produced significantly less phonation compared to pauses in both monologic and dialogic speech longitudinally over eight months. Semantically, speakers generate the content for their speech during both planning and speaking time, and the pauses during speech maybe needed to plan new semantic content (Bygate & Samuda, 2005; De Jong & Perfetti, 2011). The process of speaking involves making moment-to-moment decisions (Segalowitz, 2000). To make sure that such real-time decisions are accurate, some control mechanisms need to be carried out during speech production to verify and evaluate "the intermediate products of information processing" (Segalowitz, 2000, p. 201). Those control mechanisms act against faster speech (Segalowitz, 2000). While producing a dialogic speech, that involves interacting and negotiating meanings with an interlocutor, the processing demands of accurately verifying and evaluating intermediate thoughts might be heavier leading to even more pauses. Without such pauses, speaking performances may not efficiently meet the demands of a communicative context (Segalowitz, 2000). This might explain why in dialogic oral production in the present study, not only lower *Phonation-Time Ratio* but also lower *Articulation Rate* (indicating lower speed fluency) was predictive of higher oral proficiency scores at time three compared to time one.

Another explanation for these findings (that lower *Phonation-Time Ratio* in monologic and dialogic speech and lower *Articulation Rate* in dialogic speech were significantly predictive of higher L2 oral proficiency over time) might be the reasoning demands of the speaking tasks. In the speaking tasks (both monologic and dialogic) of the dissertation, the participants had to conceptualize specific reasons behind their choices of living, which might have been cognitively taxing for them. Previous research found that tasks that require creativity or put demands on the conceptualization stage of speech production are perceived by L2 speakers as more difficult (Préfontaine & Kormos, 2015). Such cognitively demanding tasks may require more planning time during both the conceptualization and formulation of utterances (De Jong et al., 2012a), and pauses can be attributed to "attentional preoccupation with micro-planning” (Schmidt, 1992, p. 377). Hence, to plan appropriate semantic content and produce a high-scoring speech, the participants in the dissertation might have needed to take frequent pauses (De Jong & Perfetti, 2011), which led to lower *Phonation-Time Ratio*.

Table 5.2 and Table 5.3 provide samples of two participants’ monologic speech collected over eight months. Speech samples from these two participants were chosen for analysis because they exemplify two contrasting levels of oral proficiency. Table 5.2 shows samples from participant ‘A’ whose oral proficiency scores at each time (time one: 4.5; time three: 9.84), was below the average (average at time one=11.76; average at time three=11.53, as was reported in Table 4.5). Hence, participant ‘A’ has relatively lower oral proficiency.

Table 5.2 Samples of Monologic Speech from a Participant (A) with Lower Proficiency

	Participant A		
	Time one		Time Three
	Proficiency score: 4.5/16		Proficiency score: 9.84/16
	Task: Monologic (version C)		Task: Monologic (version E)
	Length of speech extracted: 20seconds Number of syllables per run: 3.28 (total syllables 23/total runs 7)		Length of speech extracted: 20 seconds Number of syllables per run: 8 (total syllables 64/total runs 8)
	Number of pauses: 6		Number of pauses: 7
1	|*what I do*|	1	|{*But the*
2	|*and*		Pause
	Pause		*bad way*}
	once they live		Pause
	Pause		{*the*}
	for the America|		Pause
	Pause		{*like the transportation*}
3	|*how make it*		Pause
	Pause		*if you have like car or is really like far away*|
	{*any*}		
	Pause		Pause
	anything|	2	|*and you will take* {*care*} *more of the time to go to downtown Atlanta*|
	Pause		
4	|*I said*|		Pause
		3	|*So there is some the problem*|
			Pause
		4	

	|*but I prefer to take the option number two to living*|

Note. The transcribed speech in this table reflects the verbatim transcription including the fluent runs and pauses. The participants' speech is in italics. "Pause" indicates pauses above the 0.25 seconds threshold. Notations of AS-unit analysis (Foster et al., 2001) are used to indicate AS-units (enclosed in two upright slashes,| |) and false starts and repetitions (within curly brackets). Every AS-unit is numbered.

Table 5.2 shows that as participant 'A''s oral proficiency scores develop from time one (4.5) to time three (9.84), the number of syllables produced per run also increases (from 3.28 at time one to 8 at time three). However, participant 'A' is not taking more pauses to produce those longer runs because there is only a slight increase in the number of pauses from time one (6 pauses) to time three (7 pauses). As Table 5.2 also shows, participant 'A' has lower repair fluency because there are multiple false starts and repetitions which are indicated in curly brackets in participant 'A''s speech at each time. These might indicate the lack of efficient control on elements of information processing while speaking (Segalowitz, 2000). However, producing longer runs with more pauses could have allowed for higher accuracy of the online decisions during speaking (Segalowitz, 2000) and thus, less disfluency features (e.g., false starts, repetitions). Table 5.3 provide examples for this argument with samples of monologic speech from 'B', a relatively higher proficient participant. Participant 'B''s oral proficiency score at each time (time one: 12; time three: 14.45) was above average (average at time one=11.76; average at time three=11.53, as was reported in Table 4.5). Hence, participant 'B' is labelled as a participant with relatively higher oral proficiency.

Table 5.3 Samples of Monologic Speech from a Participant (B) with Relatively Higher Oral Proficiency

<table>
<tr><th colspan="4">Participant B</th></tr>
<tr><th colspan="2">Time One</th><th colspan="2">Time Three</th></tr>
<tr><td></td><td>Proficiency score: 12/16
Task: Monologic (version D)
Length of speech extracted: 20 seconds
Number of syllables per run: 5.5 (total syllables 55/total runs 10)
Number of pauses: 9</td><td></td><td>Proficiency score: 14.45/16
Task: Monologic (version B)
Length of speech extracted: 20 seconds
Number of syllables per run: 4.84 (total syllables 63/total runs 13)
Number of pauses: 12</td></tr>
<tr><td>1</td><td>|<i>you have</i></td><td>1</td><td>|<i>So</i></td></tr>
<tr><td></td><td>Pause</td><td></td><td>Pause</td></tr>
<tr><td></td><td><i>less rules</i></td><td></td><td><i>the question is whether I would</i></td></tr>
<tr><td></td><td>Pause</td><td></td><td>Pause</td></tr>
<tr><td></td><td><i>maybe even more space</i></td><td></td><td><i>prefer to</i></td></tr>
<tr><td></td><td>Pause</td><td></td><td>Pause</td></tr>
<tr><td></td><td><i>for you because on campus you're</i></td><td></td><td><i>live</i></td></tr>
<tr><td></td><td><i>usually supposed to share</i></td><td></td><td>Pause</td></tr>
<tr><td></td><td>Pause</td><td></td><td><i>in a community with a lot of my</i></td></tr>
<tr><td></td><td><i>your room with someone</i>|</td><td></td><td>Pause</td></tr>
<tr><td></td><td>Pause</td><td></td><td><i>native</i></td></tr>
<tr><td>2</td><td>|<i>and</i></td><td></td><td>Pause</td></tr>
<tr><td></td><td>Pause</td><td></td><td><i>people from my country</i></td></tr>
<tr><td></td><td><i>living off campus you will have your</i></td><td></td><td>Pause</td></tr>
<tr><td></td><td><i>own room and your</i></td><td></td><td><i>or in just international culture</i>|</td></tr>
<tr><td></td><td>Pause</td><td></td><td>Pause</td></tr>
<tr><td></td><td><i>own bathroom</i>|</td><td>2</td><td>|<i>There are certainly</i></td></tr>
</table>

3	*|and you just need to share* Pause *kitchen* Pause *Also|*	Pause *Some* Pause *advantages in living* Pause *among* Pause *your own community|*

Note. The transcribed speech in this table reflects the verbatim transcription including the fluent runs and pauses. The participants' speech is in italics. "Pause" indicates pauses above the 0.250 seconds threshold. Notations of AS-unit analysis (Foster et al., 2001) are used to indicate AS-units (enclosed in two upright slashes, | |) and false starts and repetitions (within curly brackets {}). Every AS-unit is numbered.

As can be seen in Table 5.3, participant 'B''s number of syllables per run at time one (5.5) is lower than that produced by the lower proficient participant 'A' at time three (8). However, participant 'B' is taking more pauses (9) than participant 'A' (7) while speaking. Moreover, there is no example of false starts or repetitions in participant 'B''s speech, which also expresses concise and clear ideas. Likewise, as participant 'B' increases his oral proficiency at time three (proficiency score 14.45), the number of pauses (12) increases along with a slight decrease in the length of runs (4.84). Simultaneously, there is high repair fluency in participant 'B''s speech (indicated by the lack of false starts and repetitions), which leads to the expression of efficient and appropriate ideas for fulfilling the task goals. Thus, the present findings show that ESL speakers, with development in L2 oral proficiency over time, produced speech with lower *Phonation-Time Ratio*, which indicates less phonation compared to pauses but more efficient expression of ideas characterized by a lack of disfluency features (e.g., false starts, repetitions, self-corrections). There have also been studies in literature that found significant positive relationships between pause-frequency and higher L2 oral proficiency in ESL learners' speech. For example, Riazantseva (2001) found that high proficient ESL speakers paused more frequently in their L2 (English) than in their L1 (Russian).

In contrast, Ferrari (2012), who examined longitudinal development in four Italian as L2 learners' speech over three years, found that over time L2 learners' frequency of silent pauses per AS-unit decreased. While the present study included both silent and filled pauses in pause counts (De Jong and Perfetti, 2011; Vercellotti, 2017), Ferrari (2012) only counted frequency of silent pauses. Additionally, for pause detection, the minimum duration of silence considered in Ferrari (2012) was 0.50 seconds in contrast to the 0.25 second threshold considered in the present study (De Jong & Bosker, 2013; Kahng, 2014). These methodological differences might be a reason that in Ferrari (2012), in contrast to the findings of the present study, frequency of silent pauses decreased over time.

Additionally, while producing monologic speech is dependent on the knowledge and cognitive resources of the speakers' themselves (Michel, 2011), producing appropriate dialogic speech in the dissertation involved processing the interlocutor's arguments and producing appropriate counter arguments/reasonings to reach the task goals. Because of such higher pragmatic demands of dialogic tasks (Michel, 2011), ESL speakers who developed L2 speaking proficiency over time might have needed to produce dialogic speech with lower speed fluency (i.e., lower *Articulation Rate*).

Another explanation of the high proficiency ESL speakers' lower *Phonation-Time Ratio* (in both monologic and dialogic tasks) and lower *Articulation Rate* (in dialogic tasks) over time might be the fact that during the data collection at time three, a nationwide lockdown was going on (because of the COVID-19 pandemic). Hence, at time three, the participants might have had less chance of oral interaction with others and thus, less chance of practicing speaking skills in English, which might have negatively affected their L2 oral proficiency. It might also be a possibility that the participants had reduced fluency at time three because data at time three were collected by online video meetings in contrast to the face-to-face mode of data collection at time one and time two.

5.3 Research questions 2a and 2b: The relationships between ESL speakers' WM and aptitude and their oral proficiency over time

The results of the research question 2a (Whether ESL speakers' WM and aptitude measures predict their oral proficiency scores) showed that *Explicit Aptitude* was the only cognitive variable having a significant correlation with the oral proficiency scores. However, *Explicit Aptitude* was not significantly predictive of the participants' L2 oral proficiency. Additionally, the results of the research question 2b (whether the relationships between the WM and aptitude measures and the oral proficiency scores change over time) showed that time did not have any significant interactions with the *Explicit Aptitude* scores in their effects on L2 oral proficiency. Thus, the relationships between ESL speakers' individual difference variables (e.g., *Explicit Aptitude)* and their L2 oral proficiency did not significantly vary over time. There might be several explanations for these findings.

First, the effects of aptitude on L2 learning might depend on the stages of learners' L2 acquisition (Robinson, 2007). Traditional aptitude consists of abilities (e.g., analyzing unfamiliar sounds for retention [phonetic coding], understanding the functions of words in sentences[grammatical sensitivity]) that might be important for L2 learning at initial stages, not at advanced levels (Li, 2015, 2016; Robinson, 2001; Skehan, 2012). In the dissertation, the oral proficiency levels of most of the participants ranged from intermediate to advanced, which might explain the lack of any significant relationships between *Explicit Aptitude* and the participants' oral proficiency. Winke (2013) argued that for advanced L2 learners, cognitive variables (e.g., aptitude) maybe less important than the learners' actions in social environments or the amount of time spent practicing the L2 outside of class.

Additionally, there have been arguments in literature (Li, 2015, 2016; Robinson, 2001; Skehan, 2012) that traditional aptitude tests, for example, MLAT and LLAMA, mostly tap into the ability to learn the formal or discrete aspects of a language instead of the pragmatic or contextual aspects. Previous studies (e.g., Granena, 2018; Sparks et al., 1998; Sparks et al., 2011) that found significant relationships between L2 aptitude and L2 oral proficiency operationalized oral proficiency in terms of linguistic features, for example, as CAF measures of oral production in Granena (2018) and as pronunciation, vocabulary, grammar, comprehensibility, and listening comprehension skills in Sparks et al. (1998) and Sparks et al. (2011). Previous studies examining the relationships between EWM or PM and L2 oral skills also focused on discrete linguistic aspects (for example, oral fluency in O'Brien et al., 2007; CAF measures in Ahmadian, 2012). The present study analyzed how ESL speakers' WM and aptitude variables were related to their oral proficiency, which was composed not only of the participants' TOEFL iBT speaking scores but also of their communicative adequacy scores. The rubric of communicative adequacy only focused on the functional and pragmatic aspects of oral performances (free of any linguistic features). It might be that ESL speakers' individual difference variables (e.g., *Explicit Aptitude*) are more strongly related to the linguistic aspects of their oral performances rather than the pragmatic aspects. Hence, in the dissertation, none of the WM measures were significantly related to the oral proficiency scores, and *Explicit Aptitude* was also not significantly predictive of L2 oral proficiency over time.

5.4 Research question 3: Mediating effects of L2 speakers' WM/aptitude on the relationships between the CAF measures and L2 oral proficiency

The results of research question 3 (whether the relationships between the CAF measures and the oral proficiency scores vary depending on the participants' WM/aptitude abilities) showed that for the monologic tasks, *Explicit Aptitude* had a significant positive interaction with *MRC Familiarity All Words* in their effects on L2 oral proficiency. Thus, for the participants with higher *Explicit Aptitude*, the use of more familiar vocabulary in the monologic tasks was significantly predictive of higher oral proficiency scores. In the monologic tasks in the dissertation, the participants had to choose a place of living and provide reasons for their choice. It is unlikely that for talking about the topic of choosing a housing, the participants would have to use words that native speakers of English would find unfamiliar. There are L2 studies (e.g., Crossley & Skalicky, 2019; Salsbury et al., 2011) where learners of higher oral proficiency (measured by TOEFL tests in Salsbury et al., 2011 and by ACT college placement tests in Crossley & Skalicky, 2019) did not use less familiar words in interactive speaking tasks. Crossley and Skalicky (2019) argued that the use of familiar vocabulary might not be influenced by the users' proficiency when familiar vocabulary is important in "shaping meaning" of utterances (p. 399). In the present study, the use of unfamiliar vocabulary was not necessary to fulfill the task goals efficiently because all the monologic tasks were on the same topic of choosing a living place, which was related to the speakers' daily life, and words related to daily life have higher familiarity scores (Tanaka-Ishii & Terada, 2011). In Tanaka-Ishii and Terada (2011), highly familiar words used in daily communication also correlated strongly and positively with frequency scores in spoken corpora. Additionally, in the monologic tasks, unlike the dialogic tasks, the participants' speeches were not affected by the interactive features of discourse (e.g., responding to interlocutors' questions or arguments, checking comprehension etc.) (Gass & Mackey, 2015). Hence, for fulfilling the goals in monologic speaking tasks on a familiar topic related to daily life (e.g., choosing a place to live), ESL speakers with higher analytical and logical thinking ability (i.e., *Explicit Aptitude*) might have been more likely to use familiar vocabulary that were more appropriate for fulfilling the communicative purposes of the tasks than less familiar vocabulary. The functional effectiveness of such familiar vocabulary use in monologic tasks might have led to higher oral proficiency scores. Previous studies (e.g., Crossley et al., 2019) also found that as ESL learners develop proficiency over time, they tend to use more frequent (thus, more familiar) words in monologic speech. Spoken language is used for communication, not to impress people with unfamiliar words. Thus, ESL speakers with higher *Explicit Aptitude* might norm more closely to native speaker standards, and therefore, they might not use many unfamiliar or overcomplicated vocabulary in monologic speech.

Likewise, the interaction model on the dialogic tasks showed that for the participants with higher *Explicit Aptitude*, for each increase in their *Mean Length of Fluent Runs* and *Mean Length of Clause*, their oral proficiency scores significantly decreased. Thus, the ESL speakers with higher *Explicit Aptitude* and higher oral proficiency used shorter fluent runs and shorter clauses in their dialogic oral production. In dialogic speech, the participants needed to successfully communicate with their interlocutors to reach the task goals. It might be that the ESL speakers with higher *Explicit Aptitude* and higher oral proficiency normed to native speaker standard and used shorter runs and shorter clauses in dialogic speech than unnecessarily overcomplicated language.

Furthermore, in dialogic speech, the participants interacted with their partners to fulfil the task goals. In the model of L2 speech production, based on Levelt (1999) and De Bot (1992), the process of speaking starts with L2 speakers generating what to say, which is known as macroplanning. While responding to interlocutors (e.g., addressing the interlocutors' questions and arguments), L2 speakers may need to spend longer time on macroplanning (Segalowitz, 2010) in contrast to monologic speech where no interaction with an interlocutor is required. As dialogic speech requires more macroplanning, it consumes more processing resources (Segalowitz, 2010). L2 speakers with higher *Explicit Aptitude* have higher ability of inductive learning and conscious identification of patterns in data (Granena, 2016, 2018), which can make them more sensitive to interlocutors' input in dialogic interactions (Skehan, 2019). That input is then available for subsequent and deeper processing (Skehan, 2019) involving conscious integration of the interlocutor's speech with one's own thoughts and production of appropriate output (Segalowitz, 2010; Skehan, 2009). L2 speakers with higher inductive reasoning ability (i.e., higher *Explicit Aptitude*) might carry out such processing tasks more skillfully and thus, engage in frequent interactions with their partners with higher accuracy and appropriateness. Although such frequent interactions may lead to the use of shorter turns (Robinson, 2001), such interactions may also result in higher oral proficiency scores by fulfilling the task-goals efficiently. On the contrary, the participants with lower *Explicit Aptitude* might have difficulty handling the higher need of macroplanning and the related processing demands of dialogic tasks. Such difficulties might lead to less interactions with interlocutors and higher number of disfluencies in speech (for example, repetitions, false starts etc.) (Segalowitz, 2010), which might result in lower proficiency scores.

Table 5.4 features samples of dialogic speech from two participants at time one. Speech samples from these two participants were selected for analysis because they represent two different *Explicit Aptitude* abilities (lower and higher), and they also represent two contrasting levels of oral proficiency (lower and higher). One participant, ‘C’, has relatively lower oral proficiency score, 9.66, and the other participant, ‘D’, has relatively higher oral proficiency score, 13 (considering the average oral proficiency score at time one, 11.76, as was reported in Table 4.5). Additionally, participant ‘C’ has relatively lower *Explicit Aptitude* score, 35, and participant ‘D’ has relatively higher aptitude score, 95 (considering the mean *Explicit Aptitude* score of 64, as was reported in Table 4.17).

Table 5.4 Samples of Dialogic Speech from a Lower Proficient Participant (‘C’) and a Higher Proficient Participant (‘D’)

Participant: C	Participant: D
Oral proficiency score: 9.66/16	Oral proficiency score: 13/16
Average Explicit Aptitude: 35	Average Explicit Aptitude: 95
Task: Dialogic (version: A)	Task: Dialogic (version: A)
Length of speech extracted: from 1 second to 55 seconds (including the first 35 seconds of partner’s speech)	Length of speech extracted: from 2 second to 42 seconds (including 20 seconds of partner’s speech)
Length of ‘C’’s speech transcribed here: 20 seconds	Length of ‘D’’s speech transcribed here: 20 seconds
Number of syllables per run: 6.3 (total syllables 63/total runs 10)	Number of syllables per run: 4.56 (total syllables 41/total runs 9)
Number of fluent runs: 10	Number of fluent runs: 9
Number of clauses: 10	Number of clauses: 8
Number of words per clause (excluding repetitions, false starts,	Number of words per clause (excluding repetitions, false starts, self-corrections): 4 (total words 32/total clause 8)

	self-corrections): 5.1 (total words 51/total clause 10)		
	Partner:….	1	**D**: |*I need ::to go first?*|
1	**C**: |*yes, for me, I think ::it is going to*|		**Partner**:…
	Pause	2	**D**: |*I decide ::to {live with}*
2	|*when you share a friend in the room*|		Pause
	Pause		*live in a small apartment*|
3	|*Then*		**Partner**:…
	Pause	3	**D**: |*less*|
	{then} it could be ::at first you {you}		**Partner**:…..
	have ::to go ::to have a good	4	**D**: |*yes, so for*
	relationship|		Pause
	Pause		*me, I like it*|
4	|*and {when}*		Pause
	Pause	5	|*cause now I*
	Like		Pause
	Pause		*{I} am living with*
	you know		Pause
	Pause		*other two roommates and like*|
	It is not really specific you know		
	Pause		
	your rent|		
	Pause		
5	| *you live {in the} ::when you feel*		
	bad|*...*		

Note. In the transcribed speech in this table, the partners' speeches are omitted. The transcribed speech reflects the verbatim transcription including the fluent runs and pauses. The participants' speech is in italics. "Pause" indicates pauses above the 0.25 seconds threshold. In this table, notations of AS-unit analysis (Foster et al., 2001) are used to indicate AS-units (enclosed in two upright slashes, | |) and clauses (divided by two double colons, ::). False starts, self-corrections, and repetitions are within curly brackets, {}. Every AS-unit is numbered.

As can be seen in Table 5.4, in the speech of the lower proficient participant 'C', there are multiple examples of false starts (indicated in curly brackets in Table 5.4), for example, "*when*" (in AS-unit 4), "*in the*" (in AS-unit 5). There are also repetitions in participant 'C''s speech (indicated in curly brackets in Table 5.4), for example, "*then*", "*you*" [both in AS unit 3]). In contrast, in the higher proficient participant 'D''s speech, there is only one self-correction ("*live with*" in AS-unit 2) and one repetition ("*I*" in AS-unit 5). Therefore, although participant

‘C’ is producing longer runs (6.3 syllables per run) and longer clauses (5.1 words per clause) than participant ‘D’ (4.56 syllables per run and 4 words per clause, respectively), there are more disfluency features (such as multiple false starts and repetitions) in participant ‘C’’s speech. In contrast, although the higher proficient participant ‘D’’s speech contains smaller runs and smaller clauses, those have less disfluency features and express more concise and appropriate ideas than those produced by the lower proficient participant ‘C’. Additionally, in the higher proficient (with higher *Explicit Aptitude*) participant ‘D’’s speech, there are multiple interactions with the partner, for example, asking a question (e.g., “*I need to go first*?” [AS-unit 1]) and answering a question (e.g., “*less*” [AS-unit 3]), and during interactions, speech units are usually short (Robinson, 2001). However, in the lower proficient (with lower *Explicit Aptitude*) participant ‘C’’s speech, there is no interaction with the partner whose speech occupies the first 35 seconds, and then participant ‘C’ starts speaking. Therefore, in the dialogic tasks, stronger inductive ability of the participants with higher *Explicit Aptitude* might have helped them better integrate their interlocutors’ speech with their own thoughts and produce concise, appropriate, and interactive speech leading to higher oral proficiency scores (despite such interactive speech resulting in shorter runs and shorter clauses). However, those with lower *Explicit Aptitude* might have been less successful in meeting the processing demands of the dialogic tasks. Thus, they produced less interactive speech with longer runs and clauses, that contained more disfluency features, and all of these might have led to lower proficiency scores.

Therefore, the present findings show that L2 speakers’ cognitive abilities, especially *Explicit Aptitude*, may interact with varied information processing demands of monologic and dialogic speaking tasks (Robinson, 2005b). Such interactions may affect L2 oral production features (e.g., CAF measures), which explain variances in oral proficiency scores (Robinson, 2005b).

5.5 Implications of the findings

The adjustment challenges that non-matriculated and matriculated ESL speakers face in higher education institutes in English-speaking countries, such as USA, are mostly caused by a lack of adequate English-speaking proficiency (Andrade, 2006, 2009). This fact highlights the need for developing oral proficiency of ESL speakers in academic contexts (Andrade, 2009; Benzie, 2010). The dissertation was conducted with non-matriculated and matriculated ESL speakers in an academic context. Hence, the results of the dissertation have implications for ESL speakers' oral proficiency development in relation to linguistic features of their oral production and their individual difference variables. Previous studies on L2 oral production either examined the relationship between CAF measures and L2 oral proficiency in monologic speech (i.e., Iwashita et al., 2008; Révész et al., 2016) or investigated the development of CAF measures over time (e.g., Vercellotti, 2017; Tonkyn, 2012; Ferrari, 2012). In this regard, the dissertation study breaks new ground in L2 oral production research by investigating CAF measures of both monologic and dialogic oral tasks as predictors of longitudinal development in L2 oral proficiency. Additionally, whereas previous studies mostly examined the relationships between L2 learners' individual differences in WM or aptitude and linguistic features (e.g., CAF measures) of their oral production (e.g., Ahmadian, 2012; Granena, 2018; Kormos & Sáfár, 2008; Mizera, 2006; O'Brien et al., 2007), the dissertation study is unique in its investigation into WM and aptitude as predictors of L2 oral proficiency that subsumes skills in both linguistic and pragmatic features. Furthermore, the dissertation makes a significant contribution to SLA research by examining the interactions between ESL speakers' individual differences in *Explicit Aptitude* and CAF measures of oral production in their combined effects on L2 oral proficiency.

5.5.1 Theoretical implications

While previous L2 studies found significantly higher fluency in L2 dialogic speech than in monologic speech (e.g., Ferrari, 2012; Tavakoli, 2016), the present findings showed that producing longer fluent runs in dialogic speech is also predictive of development in oral proficiency over time. Thus, this finding suggests the importance of producing longer fluent runs in interactive speech for oral proficiency development.

Additionally, previous empirical findings showed that L2 speakers with higher oral proficiency speak at a faster rate (e.g., De Jong et al., 2015; Iwashita et al., 2008). While those studies did not examine longitudinal development in L2 oral proficiency, the findings of the present study showed that over eight months, high proficiency ESL speakers produced monologic and dialogic speech with lower phonation compared to pauses. The implication is that by taking frequent pauses, L2 speakers can make more accurate online decisions in the process of speaking, which leads to efficient expression of ideas (for fulfilling the task-goals) with higher repair fluency. This finding informs theoretical understanding of what kind of linguistic features are related to longitudinal development in L2 oral proficiency. This finding implies that developing L2 oral proficiency over time involves producing concise and functionally effective speech with frequent pauses.

Additionally, the dissertation study showed that the relationships between CAF variables and L2 oral proficiency might vary depending on monologic and dialogic tasks. The findings suggest that longer clauses and less familiar vocabulary in monologic tasks but higher rate of speech in dialogic tasks were significant predictors of L2 oral proficiency. These findings imply that while analyzing linguistic predictors of L2 oral proficiency, it is theoretically important to consider what kind of speaking tasks (monologic versus dialogic) the linguistic features were elicited from. Monologic versus dialogic tasks have varying information processing demands on L2 speakers (Robinson, 2005), which, as is shown in the present findings, might variedly interact with the linguistic features of oral production in their effects on the oral proficiency scores. These findings might also inform assessment of L2 oral proficiency by highlighting the need to maintain separate benchmarks of oral proficiency for monologic and interactive oral tasks.

Furthermore, in the dissertation, ESL speakers' individual differences in *Explicit Aptitude* significantly interacted with CAF features of their monologic and dialofic oral production in predicting L2 oral proficiency. For example, the participants with higher *Explicit Aptitude* used more familiar vocabulary in monologic tasks that was predictive of higher oral proficiency scores. Theoretically, such findings offer insights into the processes underlying the production of high proficient L2 speech by indicating the importance of the speakers' *Explicit Aptitude* and the relevant linguistic predictors in the process (DeKeyser, 2012). The dissertation also found that for the participants of higher *Explicit Aptitude*, shorter fluent runs and shorter clauses in the dialogic tasks were significantly predictive of higher oral proficiency scores. Thus, the findings provide empirical support for the theoretical argument (Robinson, 2005b) that L2 learners with varied *Explicit Aptitude* might differentially respond to the information processing demands of monologic versus dialogic tasks, which might affect their production of linguistic features and oral proficiency. Such significant interaction effects also suggest how attainment of L2 oral proficiency might be influenced by different cognitive, linguistic, and contextual mechanisms.

Moreover, the dissertation analyzed the latent structure of the oral proficiency construct that has strong implications regarding the assessment of this construct. For assessing L2 oral proficiency, standardized proficiency tests (e.g., TOEFL iBT speaking, ACTFL OPI) have frequently been used in SLA literature. Based on the theoretical views on different aspects of L2 oral proficiency (Hulstijn, 2011), the dissertation employed multiple variables (TOEFL iBT speaking, communicative adequacy for monologic tasks, and communicative adequacy for dialogic tasks) to measure the proficiency construct. The output of the factor analysis showed that those multiple proficiency measures tapped into a single latent oral proficiency construct. This finding implies that communicative adequacy might be used in SLA research as a valid measure of L2 oral proficiency. Additionally, the findings of the dissertation on different linguistic predictors of L2 oral proficiency for monologic and dialogic tasks might help assessment practitioners develop rubrics for assessing L2 oral skills.

Additionally, the dissertation study examined the latent structure of the aptitude construct, measured by five tests. The results of the PCA analysis showed similar output as those of previous studies that were conducted on larger sample sizes (e.g., Granena, 2018): LLAMA E and LLAMA F significantly loaded under the same factor ("*Explicit Aptitude*" that taps into explicit inductive ability, Granena, 2018, 2019), and the SRT (serial reaction time) test scores significantly loaded under a different factor (that taps into implicit, nonanalytical, and holistic learning ability(Granena, 2016, 2018, 2019; Kaufman et al., 2010). Thus, these findings of the dissertation offer empirical support for the argument established in previous research (Granena, 2016, 2018, 2019) that there are two main components of L2 aptitude: explicit and implicit.

5.5.2 Methodological implications

While few studies examined the relationships between L2 oral production features and L2 oral proficiency over time, the dissertation adopted a longitudinal research design to examine development in L2 oral proficiency in relation to CAF measures of both monologic and dialogic oral tasks and the speakers’ WM and aptitude variables. Such a longitudinal research design offer insights into the extent to which L2 oral proficiency develops over time and the linguistic and cognitive correlates of such development. Moreover, due to the multidimensional nature of the CAF constructs, the dissertation used multiple indices to measure each of those constructs. Such a methodological choice sheds light on how distinct dimensions of the CAF constructs are related to oral proficiency development. The dissertation also used NLP-informed indices to measure lexical sophistication that tap into the depth and breadth of ESL speakers’ lexical knowledge. Furthermore, the dissertation adopted factor analytic approaches to capture the latent structures of the multicomponential constructs such as L2 oral proficiency and aptitude. Such analyses tap into the underlying structures of these constructs. Factor analysis was used to find the best composite values of L2 oral proficiency and aptitude constructs in the dissertation.

5.5.3 Pedagogical implications

As the participants in the dissertation included non-matriculated ESL learners as well as matriculated ESL speakers, the findings might offer pedagogical practitioners empirically based insights into facilitating oral proficiency development of ESL speakers. For example, the finding that lower *Phonation-Time Ratio* was predictive of longitudinal development in L2 oral proficiency might inform ESL teachers of what kind of fluency feature to emphasize for developing their learners’ oral proficiency. For developing L2 speaking proficiency, teachers might emphasize that learners practice speaking with enough pauses so that they can more efficiently produce well thought-out ideas avoiding disfluencies (e.g., false starts, repetitions).

Additionally, the present findings that high proficient ESL speakers produced longer clauses and less familiar vocabulary in monologic speech but had higher *Articulation Rate* in dialogic speech imply that teachers might need to consider the variations between monologic versus dialogic tasks while focusing on different linguistic correlates of L2 oral proficiency. For example, teachers might emphasize that for attaining higher L2 oral proficiency, ESL learners may use longer clauses and less familiar vocabulary in monologic speech but may produce faster speech in dialogic tasks. This finding might also inform L2 assessment practitioners that linguistic correlates of L2 oral proficiency differ depending on monologic versus dialogic nature of tasks, which might offer insights into assessing L2 oral proficiency.

Moreover, the present finding that individual variations in *Explicit Aptitude* may significantly interact with linguistic features of monologic and dialogic tasks in their effects on L2 oral proficiency might help pedagogical practitioners better match ESL learners with appropriate materials or practice activities to improve their oral proficiency. For example, teachers might engage ESL learners of lower *Explicit Aptitude* in dialogic activities to practice producing efficient interactive speech that might include shorter runs but communicatively appropriate language. Thus, this finding has implications about providing better support and feedback to L2 learners of different learning abilities for producing more efficient speech.

5.6 Limitations and future directions

The study has several limitations that need to be recognized and addressed in future research. Firstly, due to the COVID-19 pandemic, the data collection of time three had to be shifted to online Zoom video meetings. Although all the other procedures for collecting data were the same across the three time (time one, two, three), the use of video meetings (instead of face-to-face meetings) for data collection at time three might still affect the participants' oral production. In the dissertation, there was no significant difference in the amount of speech produced between time one and time three and between time two and time three in either monologic (time one versus three: $t(59)=0.37$, $p=0.707$; time two versus three: $p=0.375$) or dialogic tasks (time one versus three: *t*(59)=0.609, *p*=0.544; time two versus three: *t*(59)=0.302, *p*=0.763), as shown in the results of paired-sample *t*-tests. However, the virtual mode of data collection at time three might still affect the CAF measures of oral production. Additionally, between time two and time three, the participants' use of English or their exposure to English might have been negatively impacted due to the pandemic. Hence, the participants of the dissertation might not have had normal spoken interactions with their classmates. Although speculative, all these possible factors need to be kept in mind while interpreting the results of the dissertation.

Secondly, in the dialogic tasks, the participants did not interact with the same partner at each time of data collection. In the dissertation study, it was not logistically possible to maintain the same pair combinations throughout the three phases of data collection for all the participants in the dialogic tasks. Hence, the interlocutors' proficiency or the participants' level of familiarity with their interlocutors might have affected the participants' oral performances. Therefore, in the present study, pair combination (see the 'Statistical Analysis' section for more details) was added as a random intercept in the LME models, and as shown in each LME output (see the "Results" section), this random intercept explained variances in the oral proficiency scores.

Thirdly, a dialogic speech may contain interactive features that are not present in a monologue, for example, interruptions, overlaps, and between-turn pauses, which might also affect a speaker's fluency in dialogic tasks (Tavakoli, 2016). The dissertation included similar CAF measures for both monologic and dialogic tasks because one of the purposes of the study was to compare the effects of monologic versus dialogic task types on the relationships between CAF measures and L2 oral proficiency. Hence, for calculating the fluency of dialogic speech, the dissertation did not include features like the unclaimed between-turn pauses and interruptions, which are available only for the dialogic speech, not for the monologic. Analyzing how the inclusion of such interactive features in dialogic fluency measures affect the relationships between fluency and L2 oral proficiency might be an area of future research.

Fourthly, another limitation was related to the measurement of the PM (phonological memory). Although previous research (e.g., O'Brien et al., 2006; O'Brien et al., 2007) found PM to be a significant predictor of L2 speaking skills, the dissertation did not find any significant correlations between the participants' PM and their oral proficiency scores. The current dissertation used digit span tests (administered in the participants' respective L1s) to measure their PM. However, differences in the length of digit-words in different languages (for example, shorter words for digits in Chinese versus longer words for digits in Arabic) might have been a confounding variable in their digit span scores. For instance, Chinese participants might have higher digit span scores than other participants because of the shorter words for digits in Chinese. The present study also conducted two sets of separate correlations to examine whether the PM scores of the participants from the same L1 had stronger correlations with the oral proficiency scores. One correlation was run only for the Chinese (n=23) participants, and another correlation, for the Arabic (n=13) participants. However, none of the correlations were statistically significant.

Fifthly, the accuracy measure ("*Error-free AS-unit per 100 words*") included in the dissertation only evaluated whether each AS-unit contained an error or not (irrespective of the number of errors made). Thus, every AS-unit with at least one error was treated the same irrespective of the frequency of errors (e.g., one error versus four errors) in the AS-unit. Future studies on accuracy in L2 oral production need to include a measure that would take such frequency of errors into account.

Sixthly, previous research on task-based language learning (e.g., Laufer & Hulstijn, 2001) argued that incorporating task-induced target features in L2 acquisition studies might positively affect L2 learning. In the present study, although the choice of specific complexity feature (*Number of Wh-Clauses per 100 Words*) was based on L2 learners' use of this feature in the pilot task-administration, the inclusion of more task-induced complexity measures might provide a more informative insights into development of complexity in L2 learners' oral production over time.

These limitations of the current dissertation study can be addressed in future research. Future research on longitudinal development in L2 oral proficiency need to make sure that the data collection modes are consistent across all the time despite its being a challenge in longitudinal studies. Although in the dissertation, the shift to online mode of data collection at time three was unforeseen and unavoidable, future studies examining longitudinal development in L2 oral proficiency should keep consistency in data collection methodology as much as possible. Future research is also needed to gain a clearer picture of how participating in oral tasks in online versus face-to-face mode affect oral production features (e.g., CAF measures) and L2 oral proficiency. Future research can examine how difference in modality (online video versus face-to-face) affects the relationships between CAF measures in both monologic and dialogic tasks and L2 oral proficiency. The findings of such research may offer insights into the relative effectiveness of different modes of communication (e.g., online versus face-to-face) for oral proficiency development in different tasks.

Additionally, in order to avoid any confounding effects related to interlocutors in dialogic tasks, future research examining longitudinal development in L2 oral proficiency in relation to CAF measures of dialogic tasks need to pair-up the participants with the same interlocutors at each time. Such a research design may remove any confounding effects related to the variability of participants. Future research on longitudinal development in L2 oral proficiency can also divide the participants in two groups where one group with be paired with the same interlocutor for each time of data collection, and another group will be paired with a different interlocutor at each time. Comparing the linguistic features of oral performances of those two groups of participants may provide a clearer picture of how the variability related to interlocutors may affect the relationship between L2 oral production features and oral proficiency development over time. Additionally, ESL speakers' fluency in a dialogic speech might be affected by features such as overlaps or interruptions (Tavakoli, 2016). Hence, future studies examining the relationships between fluency measures of dialogic tasks and L2 oral proficiency might include the interactional fluency features (for instance, the unclaimed between-turn pauses, counts of overlaps, interruptions). Such investigations would offer insights into how the interactive features of fluency in dialogic tasks are related to ESL speakers’ oral proficiency.

Furthermore, future research on longitudinal development in L2 oral proficiency need to recruit exclusively non-matriculated participants or exclusively matriculated participants so that the findings are more clearly generalizable to a specific group of population. Moreover, to examine a stronger effect of time on L2 oral proficiency development, future studies need to collect data for a longer period than eight months (preferably more than a year). Additionally, future studies examining the relationships between L2 speakers’ digit span scores and their oral proficiency development need to recruit participants from the same L1 so that their digit span test scores are more reliable representation of the construct of PM (phonological memory).

6 CONCLUSIONS

Compared to the studies on CAF measures and L2 writing proficiency, fewer studies examined the relationships between CAF measures and L2 oral proficiency (Ortega, 2012). Developing English speaking skills is crucial for non-native English speakers' acculturation to the higher education context in the USA (Andrade, 2006, 2009). In this context, the dissertation study breaks new ground in L2 oral proficiency research by investigating linguistic features of monologic and dialogic tasks as well as ESL speakers' WM and aptitude as predictors of longitudinal development in their oral proficiency. Existing studies on the relationships between CAF measures and L2 oral proficiency mostly examined monologic speech, and few studies investigated the relationships between CAF measures and L2 oral proficiency over time. Whereas the construct of proficiency is often-neglected in L2 studies (Tracy-Ventura et al., 2014), the current longitudinal dissertation project adopted a unique methodological approach by examining the underlying nature of L2 oral proficiency from multiple operationalizations of this construct. The dissertation offers insights into the linguistics features of monologic and dialogic tasks that are predictive of longitudinal development in L2 oral proficiency. Additionally, the results of the dissertation show how CAF predictors of L2 oral proficiency vary depending on monologic versus dialogic task types. The dissertation also examined ESL speakers' WM and aptitude as predictors of their oral proficiency development. Another innovative contribution of the current study to SLA research is the investigation into *Explicit Aptitude* as a mediating factor in the relationships between CAF measures of monologic and dialogic oral production and L2 oral proficiency.

Therefore, the dissertation offers informed insights into the linguistic predictors of longitudinal development in L2 oral proficiency. The findings suggest the importance of producing longer runs in dialogic speech for longitudinal development in L2 oral proficiency. Additionally, the results of the dissertation point to the importance of taking frequent pauses in monologic and dialogic speech for oral proficiency development. The dissertation study also found that in monologic speech, high proficiency ESL speakers produced longer clauses and more sophisticated vocabulary while in the dialogic tasks, they produced faster speech. Thus, there are different linguistic correlates of L2 oral proficiency for monologic versus dialogic tasks. Moreover, the findings of the dissertation indicate that ESL speakers' use of complexity and fluency features in monologic and dialogic tasks might variedly interact with their *Explicit Aptitude* abilities in their combined effects on oral proficiency scores. This output of the study provides empirical support for Housen et al.'s (2019) argument that objective features of language use (e.g., CAF measures) and subjective speaker-related variables (e.g., aptitude) may interact to determine L2 learning outcome (e.g., in this case, L2 oral proficiency).

Overall, the current dissertation project offers empirically based insights into linguistic and cognitive predictors of L2 oral proficiency development. The findings not only inform theoretical understanding of longitudinal development in L2 oral proficiency but also offer pedagogical implications for developing ESL learners' speaking skills. Hopefully, in near future, more longitudinal studies will be carried out on L2 oral proficiency development that would offer stronger implications for developing oral proficiency of ESL speakers enrolled in higher education institutions.

REFERENCES

ACTFL. (2020, November 4). *ACTFL: Language connects*. https://www.actfl.org/center-assessment-research-and-development/actfl-assessments/actfl-postsecondary-assessments/oral-proficiency-interview-opi

Adsera, A., & Pytlikova, M. (2015). The role of language in shaping international migration. *The Economic Journal*, *125*(586), 49-81.

Ahmadian, M. J. (2012). The Relationship Between Working Memory Capacity and L2 Oral Performance Under Task-Based Careful Online Planning Condition. *TESOL Quarterly*, *46*(1), 165-175.

Ahmadian, M. J. (2015). Working memory, online planning and L2 self-repair behaviour. In E. Wen, M. Borges Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 160-174). Bristol, UK: Multilingual Matters.

American Council on the Teaching of Foreign Languages. (1999). *ACTFL proficiency guidelines–speaking: Revised 1999*. Hastings-on-Hudson, NY: ACTFL Materials Center.

Andrade, M. S. (2006). International studsents in English-speaking universities adjustment factors. *Journal of Research in International education*, *5*(2), 131-154.

Andrade, M. S. (2009). The effects of English language proficiency on adjustment to university life. *International Multilingual Research Journal*, *3*(1), 16-34.

Audacity (2019). Audacity (version 2.3.1) [Software]. Available from https://www.audacityteam.org/download/windows/

Baddeley, A.D. (1983). Working memory. *Philosophical Transactions of the Royal Society of London, 302*, 311-324.

Baddeley, A. D. (1986). *Working memory*. Oxford: Oxford University Press.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417-423.

Baddeley, A. D. (2001). Is working memory still working? *American Psychologist, 56*, 851–864.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*(3), 189-208.

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

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual Review of Psychology*, *63*, 1-29.

Baddeley, A. (2015). Working memory in second language learning. In E. Wen, M. Borges Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 17-28). Bristol, UK: Multilingual Matters.

Baddeley, A.D. & Hitch, G.J. (1974) Working memory. In G. A. Bower (Ed.), *The Psychology of Learning and Motivation* (pp. 47–89). Cambridge, MA: Academic Press.

Baddeley, A., & Logie, R.H. (1999). Working memory: the multiple component model. In A.Miyake& P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28-61). Cambridge, UK: Cambridge University Press.

Bachman, L., & Palmer, A. (1985). *Basic concerns in language test validation*. Reading, Mass: Addison-Wesley.

Balota, D.A., Yap, M.J., Hutchison, K.A., Cortese, M.J., Kessler, B., Loftis, B., … & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39,* 445-459.

Bardel, C., Gudmundson, A., & Lindqvist, C. (2012). Aspects of lexical sophistication in advanced learners' oral production. *Studies in Second Language Acquisition*, *34*(2), 269-290.

Bécue-Bertaut, M., & Pagès, J. (2008). Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics & Data Analysis*, *52*(6), 3255-3268.

Bei, X. (2010). *The effects of topic familiarity and strategic planning in topic-based task performance at different proficiency levels*. PhD Thesis, Chinese University of Hong Kong, China.

Benzie, H. J. (2010). Graduating as a 'native speaker': International students and English language proficiency in higher education. *Higher Education Research & Development*, *29*(4), 447-459.

Boersma, Paul & Weenink, David (2018). Praat: Doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 3 February 2018 from http://www.praat.org/

Bokander, L., & Bylund, E. (2020). Probing the internal validity of the LLAMA language aptitude tests. *Language Learning*, *70*(1), 11-47.

Bowden, H. W. (2016). Assessing second-language oral proficiency for research: The Spanish Elicited Imitation Task. *Studies in Second Language Acquisition*, *38*(4), 647-675.

Breheny, P. (2020, August 14). *Visreg 2.7.0.1 : An R package for the visualization of regression models.* https://pbreheny.github.io/visreg/cross.html

Bulté, B., & Housen, A. (2012). Defining and operationalizing L2 complexity. In A. Housen, I. Vedder, & F. Kuiken (Eds.), *Dimensions of L2 performance and proficiency: Complexity,* *accuracy, and fluency in SLA* (pp. 21-46). Amsterdam: John Benjamins Publishing Company.

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

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, *1*(1), 1-47.

Carroll, J. (1981). Twenty-five years of research on foreign language aptitude. In K. Diller, (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83 – 118). Rowley, MA: Newbury House.

Carroll, J . (1990). Cognitive abilities in foreign language aptitude: Then and now. In T. Parry & C. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 11 – 29). Englewood Cliffs, NJ : Prentice Hall.

Carroll, J. B., & Sapon, S. M. (1959). *Modern language aptitude test*. Psychological Corporation.

Chiswick, B. R., & Miller, P. W. (2005). Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of Multilingual and Multicultural Development 26*(1), 1-11.

Conklin, K. & Schmitt, N. (2008). Formulaic sequences: are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29* (1): 72–89.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769-786.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, *33*(4), 497-505.

Cox, T., & Davies, R. S. (2012). Using automatic speech recognition technology with elicited oral response testing. *Calico Journal*, *29*(4), 601.

Crossley, S., & Salsbury, T. L. (2011). The development of lexical bundle accuracy and production in English second language speakers. *International Review of Applied Linguistics in Language Teaching*, *49*(1), 1-26.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 243-263.

Crossley, S., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning and Technology, 17* (2), 171-192.

Crossley, S. A, Salsbury T, and McNamara D. (2009) Measuring L2 lexical growth using hyper-nymic relationships. *Language Learning* 59: 307–34.

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2014). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*(5), 570-590.

Crossley, S. A., & Skalicky, S. (2019). Examining lexical development in second language learners: An approximate replication of Salsbury, Crossley & McNamara (2011). *Language teaching*, *52*(3), 385-405.

Crossley, S. A., Skalicky, S., Kyle, K., & Monteiro, K. (2018). Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition*, 1-24.

Crowther, D. J. (2018). *Linguistic measures of second language speech: Moving from monologic to interactive speech*. [Doctoral dissertation]. Retrieved from ProQuest dissertations publishing (10808765).

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*, 67–96.

Cummins, J. (1980). The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *TESOL Quarterly*, 175-187.

Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada. *Applied linguistics*, *2* (2), 132-149.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, *26*(3), 367-396.

Davies, M. (2008-). The Corpus of Contemporary American English (COCA): 520 million words, 1990-present. Retrieved from: https://corpus.byu.edu/coca/.

De Bot, K. (1992). A bilingual production model: Levelt's "Speaking" model adapted. *Applied Linguistics, 13*, 1–24.

De Jong, N. H., & Bosker, H. R. (2013). *Choosing a threshold for silent pauses to measure second language fluency*. Paper presented at DiSS, Stockholm.

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

De Jong, N., M. Steinel, A. Florijn, R. Schoonen, and J. Hulstijn. (2012a). The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and nonnative speakers. In A. Housen, F. Kuiken, and I. Vedder

(Eds), *Dimensions of L2 performance and proficiency: Investigating complexity, accuracy and fluency in SLA* (pp. 121-142). John Benjamins.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012b). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5-34.

DeKeyser, R. M. (2009). Cognitive-psychological processes in second language learning. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 119-138). West Sussex, UK: Wiley-Blackwell.

DeKeyser, R. (2012). Interactions between individual differences, treatments, and structures in SLA. *Language Learning, 62,* 189-200.

DeKeyser, R. and Koeth, J. (2011) Cognitive aptitudes for second language learning. In E. Hinkel (ed.) *Handbook of research in second language teaching and learning vol. 2* (pp. 395–406). Routledge: Taylor and Francis.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, *63*(2), 163-185.

Derwing, T.M., Munro, M.J., & Thomson, R.I. (2007). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, 29(3): 359-380.

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31*(4), 533-557.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, *14*(1), 20.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition.* London: Lawrence Erlbaum Associates.

Doughty, C., Campbell, S., Mislevy, M., Bunting, M., Bowles, A., & Koeth, J. (2010). Predicting near-native ability: The factor structure and reliability of Hi-LAB. In M. Prior, Y. Watanabe, & S. Lee (Eds.), *Selected proceedings of the 2008 Second Language Research Forum* (pp. 10–31). Somerville, MA: Cascadilla Proceedings Project.

Ellis, N. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 63-103). Oxford: Blackwell.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, *27*(2), 305–352.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17-44.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford, UK: Oxford University Press.

Ellis, R. (2008). *The Study of Second Language Acquisition* (2nd edition). Oxford University Press.

Ellis, R. (2012). *Language teaching research and language pedagogy*. West Sussex, UK: Wiley-Blackwell.

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In A. Miyake & P. Shah (Eds.), *Models of working* *memory: Mechanisms of active maintenance and executive control* (pp. 102–134). Cambridge, UK: Cambridge University Press.

ETS. (2016). *Official TOEFL iBT tests, Volume 2.* New York: McGraw Hill Education.

ETS TOEFL. (2010). *TOEFL ibt research insight: TOEFL ibt test framework and test development.* https://www.ets.org/s/toefl/pdf/toefl_ibt_research_insight.pdf

ETS TOEFL. (2017a). *TOEFL ibt test content.* https://www.ets.org/toefl/ibt/about/content/

ETS TOEFL. (2017b). *TOEFL iBT: Test preparation*. https://www.ets.org/toefl/ibt/prepare/

ETS. TOEFL iBT. (2011). *TOEFL iBT quick prep.* Princeton, New Jersey: Educational Testing Service (ETS).

ETS. TOEFL iBT. (2015). *TOEFL iBT test questions.* Princeton, New Jersey: Educational Testing Service (ETS).

Farghal, M., & Hussein O. (1995). Collocations: A neglected variable in EFL. International Review of Applied Linguistics in Language Teaching 33 (4), 315- 331.

Ferris, D. (1998). Students' views of academic aural/oral skills: A comparative needs analysis. *TESOL Quarterly*, 289-318.

Ferris, D., & Tagg, T. (1996a). Academic listening/speaking tasks for ESL students: Problems, suggestions, and implications. *TESOL Quarterly*, 297-320.

Ferris, D., & Tagg, T. (1996b). Academic oral communication needs of EAP learners: What subject-matter instructors actually require. *TESOL Quarterly*, *30*(1), 31-58.

Fortkamp, M. B. M. (1999). Working memory capacity and elements of L2 speech production. *Communication and Cognition 32,* 259–295.

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2014). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, *43*(2), 226-236.

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.

Gass, S. and Lee, J. (2011) Working memory capacity, inhibitory control, and proficiency in a second language. In M. Schmid and W. Lowie (eds) *From Structure to chaos: Twenty years of modeling bilingualism. In honor of Kees de Bot* (pp. 59–84). Amsterdam: John Benjamins.

Gass, S. M., & Mackey, A. (2015). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 180-206). New York: Routledge.

Gass, S., Mackey, A., Alvarez-Torres, M. J., & Fernández-García, M. (1999). The effects of task repetition on linguistic output. *Language Learning*, *49*(4), 549-581.

Gatbonton, E. & Segalowitz, N. (1988). Creative automatization: principles for promoting fluency within a communicative framework. *TESOL Quarterly*, *I*(3): 473–492.

Gilabert, R., & Muñoz, C. (2010). Differences in attainment and performance in a foreign language: The role of working memory capacity. *International Journal of English Studies*, *10*(1), 19-42.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193-202.

Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning*, *63*(4), 665-703.

Granena, G. (2016). Cognitive aptitudes for implicit and explicit learning and information-processing styles: An individual differences study. *Applied Psycholinguistics*, *37*(3), 577-600.

Granena, G. (2018). Cognitive aptitudes and L2 speaking proficiency: Links between LLAMA and HI-LAB. *Studies in Second Language Acquisition*, 1-24. doi:10.1017/S0272263118000256.

Granena, G. (2019). Language aptitudes in L2 acquisition. In J. Schwieter & A. Benati (Eds.), *The Cambridge handbook of language learning* (pp. 390-408). Cambridge: Cambridge University Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts* (pp. 41–58). New York, NY: Academic Press.

Halleck, G. B. (1995). Assessing oral proficiency: A comparison of holistic and objective measures. *The Modern Language Journal*, *79*(2), 223-234.

Higgs, T. V. (1984). *Teaching for Proficiency, the Organizing Principle*. The ACTFL Foreign Language Education Series.

Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second Language Research*, *35*(1), 3-21.

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

Huensch, A., & Tracy-Ventura, N. (2017). Understanding second language fluency behavior: The effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycholinguistics*, *38*(4), 755-785.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*(3), 229-249.

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition*, *15*(2), 422-433.

Hunt, K. W. (1966). Recent measures in syntactic development. *Elementary English*, *43*(7), 732-739.

Husson, F., Josse, J., Le, S., & Mazet, J. (2020). Package ‘FactoMineR’. https://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf

Iwashita, N., Brown, A., McNamara, T., & O’Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24-49.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*, *44*(2), 137-166.

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.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuo-spatial memory span and reasoning. *Journal of Experimental Psychology: General*, 133, 189-217.

Kaufman, S.B., DeYoung, C.G., Gray, J.R., Jimenez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*, 321–340.

Kim, Y., Payant, C., & Pearson, P. (2015). The intersection of task-based interaction, task complexity, and working memory: L2 question development through recasts in a laboratory setting. *Studies in Second Language Acquisition*, *37*(3), 549-581.

Kim, Y., Tracy–Ventura, N., & Jung, Y. (2016). A Measure of proficiency or short-term memory? Validation of an elicited imitation test for SLA research. *The Modern Language Journal*, *100*(3), 655-673.

Kormos, J. (2000). The role of attention in monitoring second language speech production. *Language Learning*, *50*, 343–384.

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.

Kormos, J. and Trebits, A. (2011) Working memory capacity and narrative task performance. In P. Robinson (ed.) *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 267–286). Amsterdam: John Benjamins.

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? Language Testing, 31(3), 329–348. doi: 10.1177/0265532214526174

Kuiken, F., & Vedder, I. (2018). Assessing functional adequacy of L2 performance in a task-based approach. In N. Taguchi & Y. Kim (Eds.), *Task-Based Approaches to Teaching and Assessing Pragmatics (pp.* 265-285). John Benjamins.

Kyle, K., & Crossley, S. A. (2014). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757-786.

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, *102*(2), 333-349.

Kyle, K., Crossley, S., & Berger, C. (2017). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, 1-17.

Lazaraton, A., & Davis, L. (2008). A microanalytic perspective on discourse, proficiency, and identity in paired oral assessment. *Language Assessment Quarterly, 5*(4), 313-335.

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1-26.

Leaper, D. A., & Riazi, M. (2014). The influence of prompts on group oral tests. *Language Testing*, *31*(2), 177-204.

Leclercq, P. & Edmonds, A. (2014). How to assess L2 proficiency?: As overview of proficiency assessment research. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), Measuring L2 proficiency: Perspectives from SLA (pp. 3-23). Buffalo, NY: Multilingual Matters.

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

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

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

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics,36,* 385 – 408.

Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition*, *38*(4), 801-842.

Li, S. (2019). Six Decades of Language Aptitude Research: A Comprehensible and Critical Review. In Z. E. Wen, P. Skehan, A. Biedroń, S. Li, & R.L. Sparks (Eds.), *Language aptitude: Advancing theory, testing, research and practice* (p. 78-96). New York : Routledge.

Lightbown, P., Halter, R., White, J., & Horst, M. (2002). Comprehension-based learning: The limits of ‘do it yourself’. *Canadian Modern Language Review*, *58*(3), 427-464.

Lin, H. (2015). Computer-mediated communication (CMC) in L2 oral proficiency development: A meta-analysis. *ReCALL: The Journal of EUROCALL*, *27*(3), 261-287.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*(S1), 185-207.

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., ... & Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning*, *63*(3), 530-566.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, *21*(4), 861-883.

Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp.182-212). New York: Routledge.

Loewen, S., & Isbell, D. R. (2017). Pronunciation in face-to-face and audio-only synchronous computer-mediated learner interactions. *Studies in Second Language Acquisition*, *39*(2), 225-256.

Mackey, A., & Gass, S. (2005). *Second language research: Methodology and design*. Lawrence Erlbaum.

Mackey, A., & Sachs, R. (2012). Older learners in SLA research: A first look at working memory, feedback, and L2 development. *Language Learning*, *62*(3), 704-740.

Mackey, A., Philp, J., Egi, T., Fujii, A., & Tatsumi, T. (2002). Individual differences in working memory, noticing of interactional feedback and L2 development. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp.181-209). Amsterdam: Johns Benjamin Publishing Company.

Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning*, *60*(3), 501-533.

Martin, K. I., & Ellis, N. C. (2012). The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition*, *34*(3), 379-413.

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

Meara, P. (2005). *LLAMA language aptitude tests* . Swansea : Lognostics.

Michel, M. C. (2011). Effects of task complexity and interaction on L2 performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance*, *(pp.* 141-173). John Benjamins Publishing Company.

Michel, M. C. (2017). Complexity, accuracy, and fluency in L2 production. In S. Loewen, & M. Sato (Eds.). (2017). *The Routledge handbook of instructed second language acquisition* (pp. 50-68). New York: Routledge.

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *IRAL-International Review of Applied Linguistics in Language Teaching, 45*(3), 241-259.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review 63*(2), 81-97.

Miyake, A., & Friedman, N. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. Healy & L. Bourne (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339 – 364). Mahwah, NJ: Erlbaum.

Mojavezi, A., & Ahmadian, M. J. (2014). Working memory capacity and self-repair behavior in first and second language oral production. *Journal of Psycholinguistic Research*, *43*(3), 289-297.

Mizera, G. J. (2006). *Working memory and L2 oral fluency* (Unpublished doctoral dissertation). University of Pittsburgh, Pennsylvania.

Mostafa, T., & Crossley, S. A. (2020). Verb argument construction complexity indices and L2 writing quality: Effects of writing tasks and prompts. *Journal of Second Language Writing*, *49*, 1-23.

Mota, M. B. (2003). Working memory capacity and fluency, accuracy, complexity, and lexical density in L2 speech production. *Fragmentos: Revista de Língua e Literatura Estrangeiras*, *24,* 69-104.

Munro, M. J., Derwing, T. M., & Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech communication*, *52*(7-8), 626-637.

Nielson, K. B. (2014). Can planning time compensate for individual differences in working memory capacity? *Language Teaching Research*, *18*(3), 272-293.

Niwa, Y. (2000). Reasoning demands of L2 tasks and L2 narrative production: Effects of individual differences in working memory, intelligence and aptitude. Unpublished master’s dissertation, Department of English, Aoyama Gakuin University, Japan.

Norris, J.M. (2010). Understanding instructed SLA: Constructs, contexts, and consequences. Plenary address delivered at the *20th Annual Conference of the European Second Language Association*, Universita di Modena e Reggio Emilia Reggio Emilia, Italy, 1-4 September 2010.

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

Norris, J.M., & Ortega, L. (2012). Assessing learner knowledge. In S.M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 573-589). New York, NY: Routledge.

O'brien, I., Segalowitz, N., Collentine, J., & Freed, B. (2006). Phonological memory and lexical, narrative, and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics*, *27*(3), 377-402.

O'brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition*, *29*(4), 557-581.

Ockey, G. (2011). Self-consciousness and assertiveness as explanatory variables of L2 oral ability: A latent variable approach. *Language Learning, 61*(3), 968-989.

Olsthoorn, N. M., Andringa, S., & Hulstijn, J. (2014). Visual and auditory digit span performance in native and non-native speakers. *International Journal of Bilingualism, 18,* 663–678.

Ortega, L., Iwashita, N., Rabie, S., & Norris, J. M. (1999). *A multilanguage comparison of measures of syntactic complexity*. Honolulu, HI*:* University of Hawai ‘i, National Foreign Language Resource Center.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics*, *24*(4), 492-518.

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann, & B. Szmrecsanyi (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 127-155). Berlin: Walter de Gruyter GmbH & Co.

Ortega, L. (2013). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning, 63*(1), 1-24.

O’Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, *19*(3), 277-295.

O'Sullivan, B. (2014). Assessing speaking. In A. J. Kunnan (Ed.), *The companion to language assessment (pp.* 156-171). John Wiley & Sons, Inc.

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

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(1), 117-134.

Payne, S., & Ross, B. (2005). Synchronous CMC, working memory, and L2 oral proficiency development. *Language Learning and Technology, 9* (3), 35-54.

Payne, J. S., & Whitney, P. J. (2002). Developing L2 oral proficiency through synchronous CMC: Output, working memory, and interlanguage development. *CALICO Journal*, 7-32.

Phakiti, A. (2014). *Experimental research methods in language learning*. London, UK: Bloomsbury.

Polat, B., & Kim, Y. (2014). Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development. *Applied Linguistics*, *35*(2), 184-207.

Prefontaine, Y., & Kormos, J. (2015). The relationship between task difficulty and second language fluency in French: A mixed methods approach. *The Modern Language Journal, 99* (1), 96-112.

Pyun, O. C. (2003). *Effects of networked language learning: A comparison between synchronous online discussions and face-to-face discussions* [Unpublished doctoral dissertation]. The Ohio State University.

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

Reber, A. S., Walkenfeld, F. F., & Hernstadt, R. (1991). Implicit and explicit learning: Individual differences and IQ. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 888-896.

Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 585–594.

Révész, A. (2012). Working memory and the observed effectiveness of recasts on different L2 outcome measures. *Language Learning*, *62*(1), 93-132.

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

Rahimpour, M., & Yaghoubi-Notash, M. (2007). Examining gender-based variability in task-prompted, monologic L2 oral performance. *The Asian EFL Journal*, *9*(3), 156-179.

Riazantseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition*, *23*(4), 497-526.

Ringbom, H., & Jarvis, S. (2011). The importance of cross-linguistic similarity in foreign language learning. In M. Long & C. Doughty (Eds.) *The handbook of language teaching* (pp. 106–118). Malden: Wiley-Blackwell.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27-57.

Robinson, P. (2005a). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *IRAL-International Review of Applied Linguistics in Language Teaching*, *43*(1), 1-32.

Robinson, P. (2005b). Aptitude and second language acquisition. *Annual Review of Applied Linguistics*, *25*, 46-73.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *IRAL-International Review of Applied Linguistics in Language Teaching*, *45*(3), 193-213.

Robinson, P. & Ellis, N.C. (2008). Cognitive linguistics, second language acquisition and L2 instruction – Issues for research. In P. Robinson & N.C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 489–546). New York: Routledge.

Sáfár, A., & Kormos, J. (2008). Revisiting problems with foreign language aptitude. *IRAL-International Review of Applied Linguistics in Language Teaching*, *46*(2), 113-136.

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, *27*(3), 343-360.

Schmidt, R. (1992). Psychological mechanisms underlining second language fluency. *Studies in Second Language Acquisition 14* (14), 357–85.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in second language acquisition*, *19*(1), 17-36.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing*, *14*(2), 157–184. doi: 10.1177/026553229701400203.

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

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

Shin, S. Y., Lidster, R., Sabraw, S., & Yeager, R. (2016). The effects of L2 proficiency differences in pairs on idea units in a collaborative text reconstruction task. *Language Teaching Research*, *20*(3), 366-386.

Sime, D. (2008). "Because of her gesture, it's very easy to understand"- Learners' perceptions of teachers' gestures in the foreign language class. In S. G. McCafferty & G. Stam (Eds.), *Gestures: Second language acquisition and classroom research* (pp. 259-279). New York: Routledge.

Simpson, J. (2006). Differing expectations in the assessment of the speaking skills of ESOL learners. *Linguistics and Education*, *17*(1), 40-55.

Skehan, P. (1989). *Individual differences in second language learning*. London: Edward Arnold.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxfsord University Press.

Skehan, P. (2001). Task and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 167–185). Essex, UK: Pearson Education.

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

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

Skehan, P. (2012) Language aptitude. In S. Gass and A. Mackey (eds) *Routledge Handbook of Second Language Acquisition* (pp. 381–395). New York: Routledge.

Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In G. Granena, D. O. Jackson, & Y. Yilmaz (Eds.), *Cognitive Individual Differences in Second Language Processing and Acquisition* (pp. 17-40). Amsterdam: John Benjamins Publishing Company.

Skehan, P. (2019). Language aptitude implicates language and cognitive skills. In E. Wen, P. Skehan, A. Biedron, S. Li, & R. Sparks (Eds.), *Language aptitude: Advancing theory, testing, research, and practice*. London: Routledge.

Slama, R. B. (2012). A longitudinal analysis of academic English proficiency outcomes for adolescent English language learners in the United States. *Journal of Educational Psychology*, *104*(2), 265-285.

Solon, M., Park, H. I., Henderson, C., & Dehghan-Chaleshtori, M. (2019). Revisiting the Spanish elicited imitation task: a tool for assessing advanced language learners?. *Studies in Second Language Acquisition*, *41*(5), 1027-1053.

Son, Y. A. (2016). Interaction in a paired oral assessment: Revisiting the effect of proficiency. *Papers in Language Testing and Assessment*, *5*(2), 43-68.

Sparks , R. L. , Artzer , M. , Ganschow , L. , Siebenhar , D. , Plageman , M. , & Patton , J . (1998). Differences in native-language skills, foreign-language aptitude, and foreign-language grades among high, average, and lowproficiency foreign language learners: Two studies. *Language Testing*, *15*, 181 – 216

Sparks , R. L. , Patton , J. , Ganschow , L. , & Humbach , N . (2011). Subcomponents of second language aptitude and second language proficiency. *Modern Language Journal* , *95* , 253 – 273 .

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, *33*(2), 149-183.

Stevens, P. J. (2009). *Applied multivariate statistics for the social sciences*. New York, London: Routledge.

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning*, *65*(4), 860-895.

Syllable Count. (2018). *Syllable counter*: *Arczis Web Technologies.* http://www.syllablecount.com/

Tanaka-Ishii, K., & Terada, H. (2011). Word familiarity and frequency. *Studia Linguistica*, *65*(1), 96-116.

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.Lin

Tavakoli, P., & Foster, P. (2008). Task design and L2 performance. Language Learning, 58(2), 429–473.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-276). Amsterdam: John Benjamins.

Temple, L. (1997). Memory and processing modes in language learner speech production. *Communication & Cognition, 30* (1/2), 75-90.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning, 44*(2), 307-336.

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In L. Ortega & J.M. Norris (Eds.), *Synthesizing research on language learning and teaching* (pp. 279-298). Netherlands: John Benjamins Publishing.

Thorn, A. S., & Gathercole, S. E. (2001). Language differences in verbal short-term memory do not exclusively originate in the process of subvocal rehearsal. *Psychonomic Bulletin & Review, 8*, 357–364.

TOEFL iBT Test. (2014). Independent speaking rubrics. Retrieved from https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

TOEFL Resources. (2020). *All of the TOEFL changes in 2019-2020 (Updated).* https://www.toeflresources.com/changes-to-the-toefl-in-2018-and-2019/s

Tominaga, W. (2013). The development of extended turns and storytelling in the Japanese oral proficiency interview. In S. J. Ross & G. Kasper, *Assessing second language pragmatics* (pp. 220-257). London: Palgrave Macmillan.

Tonkyn, A. (2012). Measuring and perceiving changes in oral complexity, accuracy and fluency: Examining instructed learners' short-term gains. In A. Housen, I. Vedder, and F Kuiken (eds.) *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA* (pp. 221-245). Amsterdam: John Benjamins Publishing Company.

Tracy-Ventura, N., McManus, K., Norris, J. M., & Ortega, L. (2014). 'Repeat as much as you can': Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 143-166). Buffalo: Multilingual Matters.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory & Language, 28,* 127-154.

Vercellotti, M. L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics, 38*(1), 90-111.

Watanabe, Y., & Swain, M. (2008). Perception of learner proficiency: Its impact on the interaction between an ESL learner and her higher and lower proficiency partners. *Language Awareness*, *17*(2), 115-130.

Wang, C. Y. (2010) *A study comparing the effects of synchronous CMC and FTF interaction on L2 oral proficiency development for students with various working memory capacities [*Unpublished doctoral dissertation]. National Tsing Hua University.

Webber, B. (2008). Computational perspectives on discourse and dialogue. In S. Deborah, T. Deborah, & H. Heidi (eds.), *The handbook of discourse analysis* (pp. 798–817). London: Routledge.

Wen, Z. (2015). Working memory in second language acquisition and processing: The Phonological/Executive Model. In E. Wen, M. Borges Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing* (pp. 41-62). Bristol, UK: Multilingual Matters.

Wen, Z. (2016a). *Working memory and second language learning: Towards an integrated approach.* Bristol, UK: Multilingual Matters.

Wen, Z. (2016b). Phonological and executive working memory in L2 task-based speech planning and performance. *The Language Learning Journal*, *44*(4), 418-435.

Weissheimer, J., & Mota, M. B. (2009). Individual differences in working memory capacity and the development of L2 speech production. *Issues in Applied Linguistics*, *17*(2), 93-112.

Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 1047–1060.

Winke, P. (2013). An investigation into second language aptitude for advanced Chinese language learning. *The Modern Language Journal*, *97*(1), 109-130.

Woltz, D. J. (2003). Implicit cognitive processes as aptitudes for learning. *Educational Psychologist, 38*, 95–104.

Wood, D. (2006). Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *Canadian Modern Language Review, 63* (1): 13–33.

Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, *46*(4), 680-704.

York, J. (2019). *Language learning in complex virtual worlds: Effects of modality and task complexity on oral performance between virtual world and face-to-face tasks* [Unpublished doctoral dissertation]. University of Leicester.

Young, R. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 426-443). New York: Routledge.

Zoom Video Communications. (2020). *Zoom* [Computer software]. https://zoom.us/download

123Apps. (2020, January 3). *123Apps: Audio-converter*. https://online-audio-converter.com/

APPENDICES

Appendix A

Appendix A.1

Version A Monologic task

<u>Topic: Deciding between living alone in a small apartment versus living with roommates in a bigger apartment</u>

As an international student at Georgia State University (GSU), you are looking for housing, and you have the following two options for living:

Living alone in a small apartment

OR Living with roommates in a bigger apartment

Which way of living do you prefer? Provide reasons in detail.

You can make some notes to help you if you want.
You will have 1 minute to plan for your speech.
Your response should be between 1-2 minutes.

Appendix A.2

Version A Dialogic task

Topic: Deciding between living with only your partner in a small apartment versus living with few other roommates and your partner in a bigger apartment

Scenario: As international students at Georgia State University (GSU), you and your partner are looking for a housing to share and you have two options:

Living in a small apartment with only your partner

OR Living in a bigger apartment with your partner and few other roommates

Think about where you would like to share an apartment with your partner and why. You can make some notes to help you if you want.

During the task, you can follow the steps below:

1. Discuss with your partner where you would like to share an apartment with him/her and why.
2. Make sure to ask follow-up questions to each other to know detailed information.
3. **If both of you make the same choice**, **discuss why that choice will be good for both of you.**
4. **If you make different choices, convince your partner why your choice is better and come to an agreement.**
5. At the end of the task, **you must choose one option** that both of you agree about.

You have 1 minute to plan for your discussion.

You can do your discussion for about 2-4 minutes.

Appendix B

Appendix B.1

Version B Monologic Task

Topic: Deciding between living in an area where many people from your native country live versus in an area where many people from other countries live

As an international student at Georgia State University (GSU), you are looking for housing, and you have the following two options for living:

Living in an area where many people from your native country live

Or Living in an area where many people from other countries live

Which way of living do you prefer? Provide reasons in detail.

You can make some notes to help you if you want.

You will have 1 minute to plan for your speech.

Your response should be between 1-2 minutes.

Appendix B.2

Version B Dialogic Task

Topic: Deciding between living in an area where many people from your native country live versus in an area where many people from other countries live

Scenario: As international students at Georgia State University (GSU), you and your partner are looking for a housing to share and you have two options:

Living in an area where many people from your native countries live

Or Living in an area where many people from other countries live

Think about where you would like to share an apartment with your partner and why. You can make some notes to help you if you want.

During the task, you can follow the steps below:

1. Discuss with your partner where you would like to share an apartment with him/her and why.
2. Make sure to ask follow-up questions to each other to know detailed information.
3. **If both of you make the same choice**, **discuss** **why that choice will be good for both of you.**
4. **If you make different choices, convince your partner why your choice is better and come to an agreement.**
5. At the end of the task, **you must choose one option** that both of you agree about.

You have 1 minute to plan for your discussion.

You can do your discussion for about 2-4 minute

Appendix C

Appendix C.1

Version C Monologic Task

Topic: Deciding between living in an apartment that has high-speed internet connection and charges extra money versus living in an apartment that has no high-speed internet connection and charges no extra money.

As an international student at Georgia State University (GSU), you are looking for housing, and you have the following two options for living:

Living in an apartment that has high-speed internet connection and charges extra money

Or Living in an apartment that has no high-speed internet connection and charges no extra money

Which way of living do you prefer? Provide reasons in detail.

You can make some notes to help you if you want.

You will have 1 minute to plan for your speech.

Your response should be between 1-2 minutes.

Appendix C.2

Version C Dialogic Task

Topic: Deciding between living in an apartment that has high-speed internet connection and charges extra money versus living in an apartment that has no high-speed internet connection and charges no extra money.

Scenario: As international students at Georgia State University (GSU), you and your partner are looking for a housing to share and you have two options:

Living in an apartment that has high-speed internet connection and charges extra money

Or Living in an apartment that has no high-speed internet connection and charges no extra money

Think about where you would like to share a place with your partner and why. You can make some notes to help you if you want.

During the task, you can follow the steps below:

1. Discuss with your partner where you would like to share a place with him/her and why.
2. Make sure to ask follow-up questions to each other to know detailed information.
3. **If both of you make the same choice**, **discuss** **why that choice will be good for both of you.**
4. **If you make different choices, convince your partner why your choice is better and come to an agreement.**
5. At the end of the task, **you must choose one option** that both of you agree about.

You have 1 minute to plan for your discussion.

You can do your discussion for about 2-4 minute

Appendix D

Appendix D.1

Version D Monologic Task

Topic: Deciding between living on-campus versus living off-campus
As an international student in the USA, you have the following two options for living:

Living on campus in a student housing

or Living outside of campus renting an apartment.

Which way of living do you prefer? Provide reasons in detail.

You can make some notes to help you if you want.

You will have 1 minute to plan for your speech.

Your response should be between 1-2 minutes.

Appendix D.2

Version D Dialogic Task

Topic: Deciding between living on-campus versus living off-campus

Scenario: As international students at Georgia State University (GSU), you and your partner are looking for a housing to share and you have two options:

You can either live on-campus in a university housing

OR you can rent an apartment outside the campus.

Think about where you would like to share an apartment with your partner and why. You can make some notes to help you if you want.

During the task, you can follow the steps below:

1. Discuss with your partner where you would like to share a place with him/her and why.
2. Make sure to ask follow-up questions to each other to know detailed information.
3. **If both of you make the same choice**, **discuss** **why that choice will be good for both of you.**
4. **If you make different choices, convince your partner why your choice is better and come to an agreement.**
5. At the end of the task, **you must choose one option** that both of you agree about.

You have 1 minute to plan for your discussion.

You can do your discussion for about 2-4 minutes.

Appendix E

Appendix E.1

Version E Monologic Task

<u>Topic: Deciding between living in Atlanta downtown versus living in a residential area</u>

As an international student at Georgia State University (GSU), you have the following two options for living:

Living in Atlanta downtown

OR Living in a residential area in a city near Atlanta.

Which way of living do you prefer? Provide reasons in detail.

You can make some notes to help you if you want.

You will have 1 minute to plan for your speech.

Your response should be between 1-2 minutes.

Appendix E.2

Version E Dialogic Task

<u>Topic: Deciding between living in Atlanta downtown versus living in a residential area</u>

Scenario: As international students at Georgia State University (GSU), you and your partner are looking for a housing to share and you have two options:

Either living in Atlanta downtown

OR living in a residential area in a city near Atlanta.

Think about where you would like to share an apartment with your partner and why. You can make some notes to help you if you want.

During the task, you can follow the steps below:

1. Discuss with your partner where you would like to share an apartment with him/her and why.
2. Make sure to ask follow-up questions to each other to know detailed information.
3. **<u>If both of you make the same choice</u>, <u>discuss</u> <u>why that choice will be good for both of you.</u>**
4. **<u>If you make different choices, convince your partner why your choice is better and come to an agreement.</u>**
5. At the end of the task, **you must choose one option** that both of you agree about.

You have <u>1 minute</u> <u>to plan</u> for your discussion.
<u>You can do your discussion for about 2-4 minutes.</u>

Appendix F

Appendix F.1

Version F Monologic Task

Topic: Deciding between living in a house versus living in an apartment complex

As an international student at Georgia State University (GSU), you have the following two options for living:

You can either live in a house

OR You can live in an apartment complex

Which way of living do you prefer? Provide reasons in detail.

You can make some notes to help you if you want.

You will have 1 minute to plan for your speech.

Your response should be between 1-2 minutes.

Appendix F.2

Version F Dialogic Task

Topic: Deciding between living in a house versus living in an apartment complex

Scenario: As international students at Georgia State University (GSU), you and your partner are looking for a housing to share and you have two options:

You can either live in a house.

OR You can live in an apartment complex.

Think about where you would like to share a place with your partner and why. You can make some notes to help you if you want.

During the task, you can follow the steps below:

1. Discuss with your partner where you would like to share a place with him/her and why.
2. Make sure to ask follow-up questions to each other to know detailed information.
3. **If both of you make the same choice**, **discuss** **why that choice will be good for both of you.**
4. **If you make different choices, convince your partner why your choice is better and come to an agreement.**
5. At the end of the task, **you must choose one option** that both of you agree about.

You have 1 minute to plan for your discussion.

You can do your discussion for about 2-4 minutes.

Appendix G

TOEFL iBT Speaking Test A

Task 1

Directions: You will now be asked to speak about a familiar topic. Give yourself 15 seconds to prepare your response. Then record yourself speaking for 45 seconds.

Listen to Track 11. 0

Talk about an important experience that you recently had. Describe what happened and explain why it was important to you.

Preparation Time: 15 seconds
Response Time: 45 seconds

Task 2

Directions: You will now be asked to give your opinion about a familiar topic. Give yourself 15 seconds to prepare your response. Then record yourself speaking for 45 seconds.

Listen to Track 12. 0

Some people think that family members are the most important influence on young adults. Others believe that friends are the most important influence. Which do you agree with? Explain why.

Preparation Time: 15 seconds
Response Time: 45 seconds

Task 3

Directions: You will now read a short passage and listen to a conversation on the same topic. You will then be asked a question about them. After you hear the question, give yourself 30 seconds to prepare your response. Then record yourself speaking for 60 seconds.

Listen to Track 13. 0

Reading Time: 45 seconds

Required Work Experience

The business studies department at State University will now require all students enrolled in its program to complete one semester of work experience in a local corporation or small business. It is felt that students will benefit from this work experience by developing leadership and organizational skills that would not normally be learned in a classroom or campus setting. Furthermore, the relationships that students establish with the company that they work for may help them to secure permanent employment with that company once they have completed the program and graduated.

Listen to Track 14. 0

The woman expresses her opinion of the university's new policy. State her opinion and explain the reasons she gives for holding that opinion.

Preparation Time: 30 seconds
Response Time: 60 seconds

Task 4

Directions: You will now read a short passage and listen to a lecture on the same topic. You will then be asked a question about them. After you hear the question, give yourself 30 seconds to prepare your response. Then record yourself speaking for 60 seconds.

Listen to Track 15. 0
Reading Time: 50 seconds

The Establishing Shot

Film directors use different types of camera shots for specific purposes. An establishing shot is an image shown briefly at the beginning of a scene, usually taken from far away, that is used to provide context for the rest of the scene. One purpose of the establishing shot is to communicate background information to the viewer, such as the setting—where and when the rest of the scene will occur. It also establishes the mood or feeling of the scene. Due to the context that the establishing shot provides, the characters and events that are shown next are better understood by the viewer.

Listen to Track 16. 0

Using the professor's example, explain what an establishing shot is and how it is used.

Preparation Time: 30 seconds
Response Time: 60 seconds

Task 5

Directions: You will now listen to part of a conversation. You will then be asked a question about it. After you hear the question, give yourself 20 seconds to prepare your response. Then record yourself speaking for 60 seconds.

Listen to Track 17.

Briefly summarize the problem the speakers are discussing. Then state which solution you would recommend. Explain the reasons for your recommendation.

Preparation Time: 20 seconds

Response Time: 60 seconds

Task 6

Directions: You will now listen to part of a lecture. You will then be asked a question about it. After you hear the question, give yourself 20 seconds to prepare your response. Then record yourself speaking for 60 seconds.

Listen to Track 18.

Using points from the lecture, explain how the passion plant and the potato plant defend themselves from insects.

Preparation Time: 20 seconds
Response Time: 60 seconds

Appendix H

The rubric for rating communicative adequacy of monologic oral tasks (Kuiken & Vedder, 2018)

Content: Is the number of information units provided in the text adequate and relevant?

1	2	3	4	5	6
The number of ideas is not at all adequate and insufficient and the ideas are unrelated to each other.	The number of ideas is scarcely adequate, the ideas lack consistency	The number of ideas is somewhat adequate, even though they are not very consistent.	The number of ideas is adequate and they are sufficiently consistent.	The number of ideas is very adequate, they are very consistent to each other.	The number of ideas is extremely adequate and they are very consistent to each other.

Task Requirements: Have the task requirements been fulfilled successfully (e.g. genre, speech acts, register)?

1	2	3	4	5	6
None of the questions and the requirements of the task have been answered.	Some (less than half) of the questions and the requirements of the task have been answered.	Approximately half of the questions and requirements of the task have been answered.	Most (more than half) of the questions and the requirements of the task have been answered.	Almost all the questions and the requirements of the task have been answered.	All the questions and the requirements of the task have been answered.

Comprehensibility: How much effort is required to understand text purpose and ideas?

1	2	3	4	5	6
The performance is not at all comprehensible. Ideas and purposes are unclearly stated and the effort of the listener to understand the test are ineffective.	The performance is scarcely comprehensible. Its purposes are not clearly stated and the listener struggles to understand the ideas of the speaker. The listener has to guess	The performance is somewhat comprehensible, some sentences are hard to understand at a first listening. A second attempt helps to clarify the purposes of the speech	The performance is comprehensible, only a few sentences are unclear but are understood, without too much effort, after a second listening.	The performance is easily comprehensible and flows smoothly. Comprehensibility is not an issue.	The performance is very easily comprehensible and highly fluent. The ideas and the purpose are clearly stated.

	most of the ideas and purposes.	and the ideas conveyed, but some doubts persist.			

Coherence and cohesion: Is the text coherent and cohesive (e.g. cohesive devices, strategies)?

1	2	3	4	5	6
The performance is not at all coherent. Unrelated progressions and coherence breaks are very common. The speaker does not use any anaphoric device. The speech is not at all cohesive. Connectives are hardly ever used and ideas are unrelated	The performance is scarcely coherent. The speaker often uses unrelated progressions; when coherence is achieved, it is often done through repetitions. Only a few anaphoric devices are used. There are some coherence breaks. The speech is not very cohesive. Ideas are not well linked by connectives, which are rarely used.	The performance is somewhat coherent. Unrelated progressions and/or repetitions are frequent. More than two sentences in a row can have the same subject (even when the subject is understood). Some anaphoric devices are used. There can be a few coherence breaks. The speech is somewhat cohesive. Some connectives are used, but they are mostly conjunctions.	The performance is coherent. Unrelated progressions are somewhat rare, but the speaker sometimes relies on repetitions to achieve coherence. A sufficient number of anaphoric devices is used. There may be some coherence breaks. The performance is cohesive. The speaker makes good use of connectives, sometimes not limiting this to conjunctions.	The performance is very coherent: when the speaker introduces a new topic, it is usually done by using connectives or connective phrases. Repetitions are very infrequent. Anaphoric devices are numerous. There are no coherence breaks. The performance is very cohesive and ideas are well-linked by adverbial and/or verbal connectives.	The speaker ensures extreme coherence by integrating new ideas in the performance with connectives or connective phrases. Anaphoric devices are used regularly. There are few incidences of unrelated progressions and no coherence breaks. The structure of the speech is extremely cohesive, thanks to a skillful use of connectives (especially linking chunks, verbal constructions and adverbials),

	often used to describe relationships between ideas.

Appendix I

The sub-scale of “communicative skills/strategies” from the “paired assessment rating rubric” of Ockey (2011)

Communication Skills/Strategies (Interaction, Confidence, Conversational awareness) (based on Ockey, 2011)

1	2	3	4	5	6
Communication skills not adequate; shows no awareness of other speakers; may speak, but not in a conversation-like way	Communication skills scarcely adequate; does not initiate interaction; produces monologue only; shows some turn-taking; may say, “I agree with you,” but not relate ideas in explanation; too nervous to interact effectively	Communication skills somewhat adequate; responds to others without long pauses to maintain interaction; shows agreement or disagreement with others' opinions	Communication skills adequate; generally confident; responds appropriately to others’ opinions; shows ability to negotiate meaning quickly and relatively naturally	Communication skills very adequate; Confident and natural; asks others to expand on views; shows how own and others’ ideas are related; interacts smoothly	Communication skills extremely adequate; very confident and natural; very skillfully shows how own and others’ ideas are related; interacts very smoothly