A COMPARATIVE STUDY OF ORAL PROFICIENCY IN DIRECT (OPI) AND SEMI-DIRECT (VOCI) TESTING MODES: MEASURES OF COMPLEXITY, ACCURACY, AND FLUENCY

By

NAWAL ALI ALZAHRANI

Bachelor of Social Sciences in English
Umm Alqura University
Makkah, Saudi Arabia
2006

Master of Social Sciences in Applied Linguistics
Umm Alqura University
Makkah, Saudi Arabia
2010

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2020

A COMPARATIVE STUDY OF ORAL PROFICIENCY IN DIRECT (OPI) AND SEMI-DIRECT (VOCI) TESTING MODES: MEASURES OF COMPLEXITY, ACCURACY, AND FLUENCY

Dissertation Approved:

Dr. Gene Halleck

Dissertation Adviser

Dr. An Cheng

Dr. Stephanie Link

Dr. Shelia Kennison

.

ACKNOWLEDGEMENTS

I want to start with thanking Allah for giving me the strength, health, and abilities to complete this dissertation.

I would like to thank my advisor, Professor Gene Halleck, for her valuable time, patient guidance, and immense knowledge. I have been extremely lucky to have an advisor who worked closely with me, responded promptly to my questions, and guided me throughout this dissertation. Professor Halleck is the one who inspired me to take an interest in the field of language testing. The topic of this dissertation came from the language testing class I took with her. Professor Halleck did not only care about my academic and research success, but she also cared about my personal life. She stood by my side during the challenging times I have been through. I am greatly indebted to all the help and support I received from you. I will miss our weekly meetings, phone calls, and Skype times.

I would like to specially record my thanks and gratitude to the committee members of this dissertation, Professor An Cheng, Professor Stephanie Link, and Professor Shelia Kennison, for their insightful feedback, their professional guidance, and continuous encouragement.

I want to thank everyone who participated in this study. I am so grateful for your time and effort regardless of your very busy schedules. You made this dissertation possible.

Throughout the journey of this dissertation, nobody has been more important to me than my family. I would like to express the deepest appreciation to my husband, my rock, and life companion, Engr. Hamed Edrees, for his love, support, and endless encouragement during my graduate program. Thank you for putting up with me during this time. My two kids, Rose and Saud, are the true inspiration in my life. You are the reason I kept going.

I owe a very important debt to my parents, who taught me to always believe in myself and follow my dreams. To my father and late mother, thank you so much for giving me the life opportunities and the experiences that made me who I am today. I am also grateful to my sisters for their prayers, positive thoughts, and endless support.

Last but not least, I want to extend my gratitude to my friends, who are scattered around the world. Thank you for your calls, texts, visits, and for being there for me when I needed a friend.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: Nawal Ali Alzahrani

Date of Degree: May 2020

Title of Study: A COMPARATIVE STUDY OF ORAL PROFICIENCY IN DIRECT (OPI) AND SEMI-DIRECT (VOCI)TESTING MODES: MEASURES OF COMPLEXITY, ACCURACY, AND FLUENCY

Major Field: ENGLISH

Abstract: This study aims at comparing oral proficiency performance at two oral proficiency testing modes, namely Oral Proficiency Interview (OPI) and Video Oral Communication Instrument (VOCI) in terms of specific measures of complexity (length of ASUs, and MS-TTR), accuracy (error-free ASUs), and fluency (frequency of filled and silent pauses). It also examines the relation between task type and CAF measures in both testing modes. It further explores the test takers' perceptions and preferences towards the direct testing mode (OPI) and the semi-direct testing mode (VOCI), and then compares those perceptions and preferences with their testing performance in terms of the CAF measures. In order to achieve the goals of this study, four instruments were used to collect the data (OPI, VOCI, online background survey, and interviews conducted in Arabic). Convenience sampling was used to recruit nine senior Saudi male students, majoring in different fields in Engineering at a South-Central University in the United States. OPIs and VOCIs responses were recorded, then manually transcribed using InqScribe software. Wilcoxon Signed Rank test reveals that while complexity measures did not show any statistically significant differences in both testing modes, accuracy (Error-free ASU) and fluency (Silent pauses) showed significant differences in the OPI and VOCI testing modes. It was also found that the narrative task impacted the MS-TTR in the VOCI testing mode and the number of silent pauses in the OPI testing mode. Participants reported a variety of positive and negative perceptions towards OPI and VOCI. This study further presents information about test takers' experiences about both tests. It was also found that participants had a higher accuracy and fluency in the OPI testing that they claimed they felt more comfortable with. The current research suggests possible empirical and practical implications and some questions for future studies.

TABLE OF CONTENTS

ChapterPage

I. INTRODUCTION 1

Overview of the Chapter 1
English in the Expanding Circle (Saudi Arabia) 3
Definitions of Key Concepts 7
Research Questions 11
Research Design 11
Summary of the Findings 12
Overview of the Chapters 12

II. REVIEW OF LITERATURE 14

Overview of the Chapter 14
Testing Oral Proficiency 14
Direct Assessment 15
OPI 15
IELTS 18
Semi-Direct Assessment 19
SOPI 19
COPI 20
OPIc 20
VOCI 22
TOEFL 27
What is an Utterance? 28
CAF Framework 29
Complexity 30
Syntactic Complexity 31
Lexical Complexity 31
Accuracy 33
Specific Measures of Accuracy 34
General Measures of Accuracy 34
Fluency 35
Measures of Utterance fluency 36
Summary of CAF Measures 37

Effects of Task Type on CAF Measures and Oral Proficiency Performance ... 40

Studies That Compared Oral Proficiency Using Direct and Semi-Direct Testing Instruments ... 41

III. METHODOLOGY ... 51

Overview of the Chapter ... 51
Participants ... 51
Instruments ... 53
VOCI ... 53
OPI ... 54
Online Background Survey ... 54
Arabic Interviews ... 55
Procedures ... 55

IV. RESULTS AND DISCUSSION ... 64

CAF MEASURES IN OPI AND VOCI ... 65

Complexity Measure ... 69
Accuracy Measure ... 73
Fluency Measure ... 79
CAF MEASURES AND TASK TYPE ... 86

PARTICIPANTS' PERCEPTIONS/PREFERENCES TOWARDS DIRECT AND SEMI-DIRECT TESTS ... 93

PARTICIPANTS' PREFERENCES AND THEIR PERFORMANCE ... 122

V. CONCLUSION ... 125

REFERENCES ... 132

APPENDICES ... 161

LIST OF TABLES

Table Page

1. Measures of Complexity 37
2. Measures of Accuracy 38
3. Measures of Fluency 39
4. Test Takers' Reactions Toward Testing Modes 48
5. Participants' background information 52
6. Inter-coder Reliability 58
7. Symbols and Their Meanings 60
8. Descriptions of the Acronyms 64
9. Descriptive Statistics of Task Type and Complexity, Accuracy, and Fluency (CAF) Measures (N=9) 67
10. Wilcoxon Signed Rank Test 68
11. Signed Rank Test 69
12. Kruskal Wallis Tests for Association within OPI (N = 34) and VOCI (N = 34) for Complexity Measure 86
13. Kruskal Wallis Tests for Association within OPI (N = 34) and VOCI (N = 34) for Accuracy Measure 87
14. Kruskal Wallis Tests for Association within OPI (N = 34) and VOCI (N = 34) for Fluency Measure 87
15. Frequency of the themes of participants' perceptions towards OPI and VOC 94

LIST OF FIGURES

Figure Page

1. VOCI screenshot showing time remaining 23

2. VOCI screenshot showing restaurant-themed 24

3. VOCI screenshot showing past-tense narration question 24

4. VOCI screenshot showing vocabulary-themed question 25

5. VOCI screenshot showing written questions on the screen 26

6. VOCI screenshot showing apology-themed question with voice-over.....27

7. Taxonomic model of L2 complexity.....32

8. MS-TTR in the VOCI Testing Mode.....88

9. SP in the OPI Testing Mode 90

10. Qualitative relation between CAF measures and testing mode 123

CHAPTER I

INTRODUCTION

Overview of the Chapter

This chapter begins with a broad introduction that shows the motives for conducting this research. It then briefly describes the norms of learning and teaching the English language in the expanding circle, more specifically in Saudi Arabian public education. Key concepts are then defined, in both a constitutive sense that shows how the term is defined in the literature and an operational sense that explains how terms were defined for the study. The research design and research questions are then described in detail, followed by a summary of the main findings of this study. This chapter closes with an overview of the remaining chapters of this dissertation.

INTRODUCTION

Thinking back about my journey in learning and later teaching English, I realize that oral proficiency is still neglected in Saudi Arabian public schools. Throughout my education there, I believe that my oral proficiency was never tested in a comprehensive manner. At the university level, I had to take listening and speaking courses, and listening comprehension tests. But speaking skill was assessed only through classroom discussion

and oral presentations, and I was never aware of any clear guidelines or framework for measuring my oral proficiency.

After my studies, when I began teaching at the English Language Institute at another Saudi university, our mandated curriculum plan ignored oral proficiency, focusing instead on grammar, vocabulary, and other aspects of language learning. I believe that the unavailability of certified language testers and the cost of hiring outside testers may have contributed to this neglect of oral proficiency testing. When I later came to the United States and met several Saudi students who had taken standardized English tests, I learned that many of them had taken both TOEFL and IELTS. Interestingly, they reported having received different speaking scores on each test and they believed that getting different scores was related to the mode of delivery, particularly the presence of a human examiner in the speaking section of the IELTS. For example, some students believe that in the IELTS test, the score on the speaking section depends on the rater. One of my study participants said:

انا اختبرت الايلتس مرتين، المره الاولي اخذت ٥ واللي اختبرني كان من باكستان او يمكن الهند، المره الثانيه اخذت ٦.٥ واللي اختبرني كان امريكي. هذا الشي خلاني احس ان اذا كان اللي يختبرك

native speaker or non-native

يفرق، لان الامريكان متسامحين اكثر في الاجابات هم يركزون على هل فهموك ولا لا، وهل جاوبت السؤال ولا لا، لكن اذا كان non-native speaker

يكونون شديدين اكثر يركزون على القرامر والكلمات اللي تختارها وحتى علي طريقة نطقك ويمكن حتى ينسى يركز عالصوره الكليه

I took IELTS twice and the first time, I got 5.00 in speaking and the rater was from Pakistan or India, and the second time I got 6.5 and the rater was American. This makes me feel that it makes a difference if the rater is native or non-native speaker because American raters are more tolerant with our answers and they see whether they understand you or not, whether you answered the question or not; however, when the raters are non-native speakers, they are more strict and they focus on your grammar, your vocabulary, and even your pronunciation, and they probably forget to look at the big picture of my answer.

The issue of oral proficiency testing in Saudi Arabia drew my attention, and I decided to investigate oral proficiency testing of Saudi Arabian learners, especially the impact of testing modes. In order to understand the background of ESL participants who were recruited in this study, I will provide a brief description of educational norms in the Saudi system.

English in the Expanding Circle (Saudi Arabia)

In Kachru's (1992) model of World Englishes, Saudi Arabia is part of the expanding circle, where English is spoken as a foreign language. According to Lowenberg (2002), the norms for English usage, teaching, and testing in this context are different from those in the inner circle, where English is the native language. Current research indicates that in some expanding circle countries, English actually functions as a second language, such as India. However, in other typical expanding circles contexts, English is used as a foreign language, such as English in Saudi Arabian context.

English teaching and learning in Saudi Arabia is teacher-centered, where the teacher dominates the class (Fareh, 2010), mainly by lecturing. As a former student in the Saudi public education system, I can confirm that this description matches my own experience, where the only speaking that the students did in class was answering short questions from the teacher. Another challenge in teaching English in Saudi Arabia is the "reliance on outdated methods" (Ahmad, 2014, p. 99), such as the grammar-translation and audio-lingual methods, which completely neglect speaking skills.

However, more recent research (Liton, 2012) has reported changes in Saudi attitudes toward learning and teaching English, due to the growing need for English in both classrooms and the workplace. Within the Middle East, Saudi Arabia has become a leading employer of English-speaking foreign professionals in its schools, hospitals, and companies (Alasmari & Khan, 2014). This development motivates Saudis to learn English and be able to communicate with the professionals globally. English has also become the medium of instruction in some undergraduate programs, such as medicine, computer science, and engineering. However, the language students learn in these programs is focused on field-specific technical terminology, with very limited communicative functions.

Alasmari and Khan (2014) claim that in Saudi public schools and universities, "EFL teachers are asked to follow the prescribed curriculum while designing teaching strategies both for classroom instructions and evaluation" (p. 318). In some preparatory year programs in Saudi universities, instructors have to follow certain guidelines and cover a given number of chapters and books. Then, at the end, all the students receive

unified tests. There is too much pressure on the instructors because they have to finish the required materials, which shifts their attention from quality to quantity. I need to emphasize that the educational sector I have been describing in this section is only related to the public schools and universities in Saudi Arabia, as there are many international schools in Saudi Arabia that use English as the main language of instruction and use teaching methodologies that are based on American or British curricula.

Students in Saudi public schools start learning English in 4th grade and continue through the end of their secondary education. English class typically lasts for an hour a day, with a heavy focus on grammar, and minimal focus on oral proficiency. By the time students reach university, they may be familiar with all the rules of English grammar, but they may struggle to speak the language with any fluency. Saudis wishing to study abroad need to take a standardized test to qualify for admission to foreign universities. The majority of students are unable to get the required exam score and enroll in specialized English language programs to improve their proficiency and prepare them to retake the standardized admission test, typically the IELTS or TOEFL. Both of these tests include a speaking component. The IELTS speaking test is conducted face-to-face and is considered direct, whereas the computer mediated TOEFL speaking test is considered semi-direct.

Little is known about the perceptions and attitudes of Saudi ESL learners towards direct and semi-direct testing modes of oral proficiency. Many students take both tests and find a variation between their speaking scores in the two exams. This study will give some insights about the differences between direct and semi-direct oral proficiency

testing. All study participants attended both Saudi public schools and an intensive English program in the United States and had taken both IELTS and TOEFL in order to be admitted to their Engineering programs at a university in the south-central United States.

Despite years of study, Saudi students of English often have limited oral proficiency (Alshumaimeri, 2003). This problem may be caused by inadequacies in the curriculum, teaching methodologies, and testing tools (Al-Nasser, 2015; Bacha, 2002; Javid, 2011; Rabbah, 2003; Tahaineh, 2010). Specific issues include an outdated focus on grammar and vocabulary (Alhmadi, 2014), inadequate time spent on speaking practice (Alhmadi, 2014; Hamad, 2013), and teacher-centered methodologies (Al Asmari, 2013; Al Hajailan, 2003; Fareh, 2010). Assessment focuses primarily on grammar, with minimal attention to listening and speaking skills (Al-Nasser, 2015; Al-Nofaile, 2010; Rahman & Alhaisoni, 2013; Alsudais, 2017; Gubaily, 2012). As often occurs in EFL contexts, Saudi oral proficiency testing often relies on invalid methods (Al Mineeai, 2013; Al-Ma'shy, 2011). For example, Alfallaj and Al-Ahdal (2017) claimed that speaking skill is sometimes assessed through written tests. Al Asmari (2013) stated that teaching methodology does not lead to learning as teachers occupy the space inside the classrooms, which makes the students more dependent on their teachers. Another factor limiting Saudi oral proficiency testing is large class sizes (Aljarf, 2006; Bahanshal, 2013; Khazaei et al., 2012; Sook, 2003), which make reliable testing methods too time consuming, especially in the absence of practical testing tools (Al Hassan, 2019; Farooqui, 2007).

Alharbi and Surur (2019) reported that "attempts to evaluate assessment techniques and procedures, especially for speaking skills, are lacking in a Saudi context" (p.1). Hosseini and Azarnoosh (2014) found that Saudi teachers rely on presentations and discussion to assess oral proficiency but were often unsure about which aspects of students' production they should assess. Several researchers (Ahmed & Alamin, 2014; Al-Seghayer, 2011; Hosseini & Azarnoosh, 2014; Noor, Muniandy, Shanmugan & Mathai, 2010) have emphasized the need for more reliable oral testing, especially through the use of technological tools.

The purpose of the present research is to investigate if I can find a reliable and valid tool for testing students' oral proficiency in the Saudi context.

Definitions of Key Concepts

VOCI (Video Oral Communication Instrument)

Constitutive definition

The English version of the VOCI refers to the multimedia-enhanced oral proficiency test developed by Halleck and Young in 1995 at San Diego State University's Language Acquisition Resource Center. It incorporates both visual and audio input presented through an audiovisual tape.

Operational definition

The term refers to the semi-direct instrument used in this study. In this study, I used the English version of the VOCI.

Face-to-face OPI (Oral Proficiency Interview)

Constitutive and operational definition

According to the Language Testing International (LTI), OPI is a standardized oral proficiency test for foreign language learners. A certified ACTFL evaluator interviews the testee, in person or by telephone. After some preliminary background questions, the interviewer adapts the rest of the interview depending on the testee's oral proficiency level.

Complexity

Constitutive definition

Complexity was defined by different researchers. For example, Ellis (2003, p. 340) referred to complexity as "the extent to which the language produced in performing a task is elaborate and varied." Later in 2009, Ellis modified his definition to refer to "the capacity to use more advanced language" (p. 475). Vercellotti (2012) also added another description of complexity, which is "the language that is at the upper limit of the student's interlanguage system, which is not fully internalized or automatized by the learner" (p. 14).

Operational definition

In this study, complexity was measured using the number and the mean length of Analysis of Speech Units (ASUs), and the Mean Segmental of Type-Token Ratio (MS-TTR). The mean length of the ASU is measured by dividing the total number of tokens, except the repeated tokens or phrases, by the number of ASUs per 100 words. The MS-

TTR is calculated by calculating the TTR for every 100 words, followed by calculating the total mean for all TTRs.

Accuracy

Constitutive definition

Accuracy is the most internally coherent construct. Pallotti (2009) defined accuracy as "the degree of conformity to certain norms" (p.4). Similarly, Hammerly (1991) and Wolfe-Quintero, Inagaki, and Kim (1998) defined accuracy as the degree of deviation from specific norms. Housen and Kuiken (2009) defined accuracy as error-free speech. This measure can be specific (e.g., accuracy of verb forms) or general (e.g., overall number of errors or error-free units).

Operational definition

Accuracy refers to the percentage of error-free ASUs per 100 words, based on errors that do not follow the prescriptive rules of English grammar.

Fluency

Constitutive definition

In the field of second language acquisition, fluency has long been a subject of general interest (e.g., Chambers 1997; Freed 2000; Guillot 1999; Hilton 2008; Lennon 1990; Koponen & Riggenbach 2000). In defining the CAF framework, Housen and Kuiken (2009) define fluency as "general language proficiency, particularly as characterized by perceptions of ease, eloquence and 'smoothness' of speech or writing" (p.4). However, there are more specific definitions of fluency. For example, Ellis and

Barkhuizen (2005) defined fluency as the production of language in real time without undue pausing or hesitation. According to Tavakoli and Skehan, (2005), fluency is a multidimensional concept that has sub-dimensions such as breakdown fluency, repair fluency, and speed fluency

Operational definition

In the current study, fluency refers to the frequency of silent pauses, and filled pauses with “ah” and “like” per 100 words.

Analysis of Speech unit (ASU)

Constitutive and operational definition

An ASU is an utterance containing “an independent clause, or sub-clausal unit, together with any sub-ordinate clause(s) associated with either” (Foster, Tonkyn, & Wigglesworth, 2000, p. 365).

Errors

Constitutive definition

Error is considered to be “any digression in syntactical, morphological, and lexical norms, but not in punctuation or capitalization” (Ruiz-Funes, 2014, p. 174).

Operational definition

Any deviations from the prescriptive norms of English grammar, especially with regard to tense, aspect, subject-verb agreement, or part of speech.

Research questions

This study examines language production in the context of oral proficiency testing. It attempts to determine whether complexity, accuracy, and fluency are affected by task type, testing mode, and attitudes toward direct and semi-direct testing modes.. The research questions of the study were:

Q1: Are there any differences in participants' complexity, accuracy, and fluency measures while taking the VOCI and OPI?

Q2: Are there any differences in the accuracy, complexity, and fluency of test takers' utterance in terms of different task types?

Q3: What are participants' perceptions of the two exams? Which testing mode do they prefer, and why?

Q4: Do participants' mode preferences impact their language production on the OPI and VOCI?

Research Design

The researcher collected diverse types of data using two testing instruments (OPI and VOCI), an online background survey, and face-to-face, semi-structured Arabic interviews. Analysis of the testing data focused on the degree to which task type and testing mode affected complexity, accuracy, and fluency. The Arabic interview was conducted in order to determine participants' perceptions and attitudes relating to direct (OPI) and semi-direct (VOCI) testing modes.

This study utilized the exploratory sequential mixed method approach (Creswell, 2014). The initial qualitative phase consisted of administration of the OPI, followed by the VOCI. Perceptions and preferences of these testing modes were then explored in the Arabic interviews. In the subsequent quantitative phase, participants' language production on the two exams was analyzed for complexity, accuracy and fluency, using descriptive and inferential statistics. Learner production on the exams was analyzed to determine whether testing mode and task type had a significant impact on performance.

Summary of the Findings

Findings of this study reveals that while complexity measures did not show any statistically significant differences in both testing modes, error-free ASU accuracy measure and silent pauses for the fluency measure showed significant differences in terms of those measures in OPI and VOCI testing modes. It was also found that the narrative task impacted the MS-TTR in the VOCI testing mode and the number of silent pauses in the OPI testing mode. Participants reported a variety of positive and negative perceptions towards OPI and VOCI. This study further presents information about test takers' experiences about both tests. It was also found that participants had a higher accuracy and fluency in the OPI testing that they claimed they felt more comfortable with. The current research suggests possible empirical and practical implications and some questions for future studies.

Overview of the Chapters

The remaining chapters of this dissertation are organized as follows. Chapter 2 presents the literature review. Testing oral proficiency is briefly discussed. Direct and

semi-direct assessment are described. CAF measures are defined and discussed in detail, followed by defining the utterance, as it is the focus of this study. The effects of task types on CAF measures will be presented. Then, studies that compared the direct and semi-direct assessment will be reviewed. The significance of the study is presented at the end of this chapter.

Chapter 3 presents the methodologies used in this research. Participants are described in detail, as are the four instruments used to collect data. The research design is presented, and the procedures used to conduct this research are described.

Chapter 4 presents the results and discussion. This chapter is divided into four subsections, starting with the descriptive statistics of the participants' complexity, accuracy, and fluency in both testing modes for each of the four tasks. After that, each CAF measure will be discussed using examples from different task types. Then, the results and discussion of the relation between task type and CAF measures will be presented. Discussion of the participants'' perceptions and preferences towards direct (OPI) and semi-direct (VOCI) testing modes will be provided. At the end of the chapter, the relation between testing modes perceptions and preferences and test performance will be discussed.

Chapter 5 will present my conclusions about the findings and discusses the implications and limitations of the research.

CHAPTER II

REVIEW of LITERATURE

Overview of the Chapter

This chapter starts with a brief description of testing oral proficiency. After that, the direct assessment will be discussed illustrating different types of direct tests. Semi-direct assessment will follow with a description of different semi-direct tests. after that, a definition of an utterance will be provided. Then, complexity, accuracy, and fluency (CAF) will be discussed in detail. Then, CAF measures will be summarized. After that, effect of task types on the CAF measures will be explained. Finally, studies that compared direct and semi-direct testing modes are reviewed.

Testing Oral Proficiency

Language proficiency is often defined as the ability to use the four skills (speaking, writing, listening, and reading), in spontaneous, authentic contexts (ACTFL, 2012, p. 3). Oral proficiency has been defined more specifically as "knowledge and automated ability for use of core vocabulary and grammar delivered with reasonably intelligible pronunciation and fluency" (Wu & Ortega, 2013, p. 681).

Oral proficiency has been tested using large-scale standardized oral proficiency tests, which are typically classified as either direct or semi-direct. Direct tests include the

Oral Proficiency Interview (OPI) and the speaking section of the International English Language Testing System (IELTS). Semi-direct tests are either tape-based or computer based and include the Simulated Oral Proficiency Interview (SOPI), Computerized Oral Proficiency Instrument (COPI), Video Oral Communication Instrument (VOCI), and the speaking section of the Test of English as a Foreign Language (TOEFL).

Direct Assessment

This section illustrates some types of direct tests of oral proficiency, including OPI, which is one of the instruments used in this study, and IELTS. Although the focus of this study is on the OPI, description of the IELTS was included as a secondary level of comparison with other semi-direct tests.

OPI

OPI has been used in the United States since the Second World War, in order to assess the language skills of American personnel working abroad. The OPI has been used extensively by the Foreign Service Institute (FSI), the Defense Language Institute (DLI), the Language School of the Central Intelligence Agency (CIA), and the United States Peace Corps. It is one of several tests used by the US government that are based on American Council on the Teaching of Foreign Languages (ACTFL) guidelines. According to Higgs (1984) "the ACTFL guidelines reflect a convergence of the governmental and academic educational sectors" (p. 22).

The OPI is a 15- to 30-minute interview that assesses the interviewee's functional ability in a second or foreign language. It can be conducted face-to-face or over the

telephone, and it includes at least one trained examiner or rater (two in the case of the FSI OPI). In 1982, language educators started using the OPI in the academic disciplines and at that time the ACTFL published its proficiency guidelines, which were last revised in 2012. The ACTFL guidelines measure proficiency across a continuum, from full professional proficiency to little or no functional ability. ACTFL tests that were designed to "evaluate speech that is either Interpersonal (interactive, two -way communication) or Presentational (one-way, non-interactive)" (ACTFL, 2012, p. 4). The guidelines show the characteristics of integrated performance in all of the four skills, based on skill descriptions used by the Interagency Language Roundtable (ILR). However, the ILR was not as concerned with distinguishing between the proficiency levels at the lower levels. On the contrary, ACTFL was not as focused at the higher levels because that was not what their subjects are generally tested (Arnett & Haglund, 2001; Liskin-Gasparro, 1984a & 1984b; Omaggio, 1986). Higgs (1984) referred to the ACTFL guidelines as "descriptive, rather than prescriptive, based on experiences rather than theory" (p.37).

ACTFL speaking guidelines were used extensively in the field of oral language testing, especially for the ACTFL Oral Proficiency Interview (OPI). In 1999, speaking guidelines were reevaluated, and the presentation of proficiency levels was changed. They are now presented in a top-down fashion (Superior to Novice) instead of bottom-up (Novice to Superior), to present a more positive evaluation of learners' performance by focusing on what they can do, rather than what they cannot do. The guidelines are divided into four main proficiency levels (superior, advanced, intermediate, and novice). Apart from the first level, the other three levels contain three sublevels, which are high, mid, and low.

When the OPI was introduced to academia, language teachers viewed it as a valid instrument for assessing learners' oral proficiency (Clark & Clfford, 1988; Dandonoli & Henning, 1990; Halleck, 1992; Kuo & Jiang, 1997; Reed & Halleck, 1997). Although the ACTFL OPI has been considered an efficient oral proficiency assessment tool (Clark & Clifford, 1988; Dandonoli & Henning, 1990; Kuo & Jiang, 1997), some scholars criticized its validity and reliability (Bachman, 1998). For example, Pienemann, Johnstone and Brindley (1988) criticized the ACTFL scale, arguing that "such descriptions are so vague and general as to be utterly unhelpful in distinguishing any second language learner from another" (p.129).

Okada (2010) argued that studies criticizing the validity of the OPI did not examine all the tasks in the OPI, as for example, very few had examined the OPI role play task. Okada said, "without investigating other activities, any claim of the validity of OPIs may not be validated sufficiently" (p. 1648).

While some researchers questioned the definitions of the ACTFL Guidelines (Bachman & Savignon, 1986; Lantolf & Frawley, 1985; Valdman, 1988), Lowe (1986) and Higgs (1984) claimed that they had been used successfully and should stand as the framework of the proficiency movement. Other scholars found an acceptable level of reliability in the ACTFL OPI (Dandonoli & Henning, 1990; Halleck, 1996; Surface & Dierdorff, 2003; Thompson, 1995, 1996). Even though some researchers criticized the OPI, Liskin-Gasparro (2003) claimed that there are three factors that make the ACTFL OPI survive and still be used extensively by policy makers, language educators, and test developers. First, it had been "a catalyst for major change in foreign language teaching at

all levels" (Liskin-Gasparro, 2003, p. 486). Second, its focus on tasks had inspired the pro-achievement tests and hybrid tests. Third, an extensive amount of proficiency research is based on OPI testing, because "OPI speech samples make for rich data" (Liskin-Gasparro, 2003, p. 488). For example, OP1 discourse has been used to compare the nature of language in face-to-face and semi-direct tests (e.g., Koike, 1998; Shohamy, 1994) and to examine how oral discourse is jointly constructed jointly by the interviewer and examinee (e.g., Ross & Berwick, 1992; Ross & Kasper, 1998). In addition, Henning (1992) claimed that the use of ACTFL guidelines in OPI rating is advantageous, making the OPI a useful assessment tool. He found that the use of ACTFL guidelines in OPI rating is advantageous, making the OPI a useful assessment tool. OPI is being used extensively in academia, especially for teacher certification (Hammadou Sullivan, 2011; Malone & Montee, 2010). OPI testing is required for aspiring foreign language teachers in 23 US states, 16 of which require an ACTFL rating of Advanced Low or higher for licensure (Kissau, 2014).

IELTS

IELTS is a standardized test that assesses the four skills (listening, reading, writing, and speaking). IELTS was originally known as ELTS (English Language Testing Service) (O'Sullivan, 2012). ELTS was developed in the United Kingdom to test English for specific purposes. Specifically, it assessed the proficiency of students who wanted to major in specific fields, such as Life Sciences, Social Studies, Physical Sciences, Technology, and Medicine. Later, when IELTS replaced ELTS, domain-specific modules were removed, and all students took the same test (O'Sullivan, 2012). There are two

versions of the IELTS, Academic and General, which differ in content, context, and purpose.

The speaking section of the IELTS is conducted face-to-face, with test takers interviewed by certified IELTS testers. This section of the test is administered by the British Council. The speaking section takes 11-14 minutes and consists of three parts. In the first part, test takers are asked about daily life issues. In the second, they receive a card with questions on it, and are given one minute to read the questions and prepare their answers. Then, they have to speak from 1-2 minutes. In the third part, testees are asked discussion questions related to their responses from the second part. Examinees' responses are recorded and sent to the British Council center for evaluation and scoring.

Semi-Direct Assessment

SOPI (Simulated Oral Proficiency Interview)

SOPI is a tape-mediated, performance-based speaking test. It uses instructions delivered by an audio-tape and a booklet in order to elicit language samples from test takers. SOPI is preferred over IELTS by many school systems and universities. It does not require certified examiners, and has other advantages related to time, location, and cost. Although the SOPI is old, Sandford university is still using this test for undergraduate students taking foreign languages.

Several studies have shown a strong positive correlation between SOPI and OPI (Clark & Li, 1986; Kenyon & Stansfield, 1993; Shohamy, Gordon, Kenyon, & Stansfield, 1989; Stansfield, 1992; Stansfield et al., 1990). Furthermore, Stansfield et al. (1990)

found that SOPI appeared to be more reliable and easier to rate than OPI. In addition, Stansfield (1996) found that SOPI is advantageous in several ways. For example, any teacher, technician or language lab can administer SOPI. Also, SOPI can be administered in locations where trained raters and ACTFL-certified testers are not available. Furthermore, SOPI can be administered individually or for a group of examinees, unlike the OPI, which can only be administered individually. However, one of the shortcomings of the SOPI is that it is not available in English.

COPI (Computerized Oral Proficiency Interview)

The Center of Applied Linguistics (CAL) adapted the SOPI to develop a computer-mediated test called the COPI. For this test, the computer stores a broad range of task-based questions, and administers them based on self-assessed proficiency level, and demographic characteristics such as age, gender, and native language. COPI can also allow examinees of high proficiency to hear the instructions in the target language (Malabonga, Kenyon, & Carpenter, 2005).

OPIc (Oral Proficiency Interview – Computer)

In 2006, ACTFL developed the OPIc, a computer-mediated version of the OPI, to overcome the OPI's reliance on in-person examiners. The OPIc is administered by Language Testing International, takes 20-40 minutes, and can be taken from any computer that is connected to the internet. It is currently available in 14 languages (Arabic, Mandarin Chinese, English, French, German, Italian, Japanese, Korean, Pashto, Persian Farsi, European Portuguese, Brazilian Portuguese, Russian, and Spanish), and is most common in South Korea. It is used extensively in the field of business, and more

sparingly in government and education. In the educational setting, its uses include language program placement, admissions, monitoring students' progress, and program evaluation.

Test questions are selected by the computer based on examinees' self-assessed proficiency level and areas of personal interest, which they provide in a pre-exam background questionnaire. ACTFL tries to limit the repetition of interest areas and prompts (Isbell & Winke, 2019). One unique feature of OPIc is that test takers can replay the prompt when the question finishes playing. Test takers are also given time to prepare for their answers (30-120 seconds).

Examinees' responses are recorded and later evaluated by one or two certified raters. The rating scale is based on ACTFL guidelines, except that it does not include the "distinguished" level. ACTFL provides extensive rater training, which "involves a four-day initial training and certification followed by regular benchmarking and norming activities" (Isbell & Winke, 2019, p. 471). This kind of training makes the OPIc a reliable testing tool. Isbell and Winke (2019) have also claimed that a key strength of OPIc is that test takers can easily understand their scores, using the available supplementary materials. However, OPIc is not without problems. Researchers have claimed that OPIc is simplistic and does not assess many aspects of oral proficiency (Isbell & Winke, 2019), and that OPIc should be rated in greater depth, especially when there is only one rater (Knoch & Chapelle, 2018).

VOCI

The English version of the VOCI is an oral proficiency instrument that was developed at San Diego State University's Language Acquisition Resource Center (Halleck & Young, 1995). The VOCI is considered a semi-direct test, as it incorporates both visual and audio input presented through an audiovisual tape, DVD, or computer file, and it does not involve a live interviewer who communicates directly with the test taker. The test taker is responding to questions delivered through the computer.

The VOCI uses carefully designed tasks to collect speaking samples that are rated based on the ACTFL scale (Kenyon, 1998). The test consists of 36 questions that assess four levels of proficiency: novice, intermediate, advanced, and superior. The first three questions are mainly for acquainting the test takers with the test and ensuring that sound and pictures are clear. This computerized test provides different situations or scenarios, which are followed by a question for test takers. This test can be administered individually, which is the case in this study, or it can be done for a group of test takers. There are two versions of this test, timed and untimed. The 45-minute timed test is the one used for this research. In this timed version, there are green bouncing balls that decrease in number to represent the remaining time for the given task (as shown in Figure 1). The test taker is supposed to finish speaking before the disappearance of the bouncing balls.

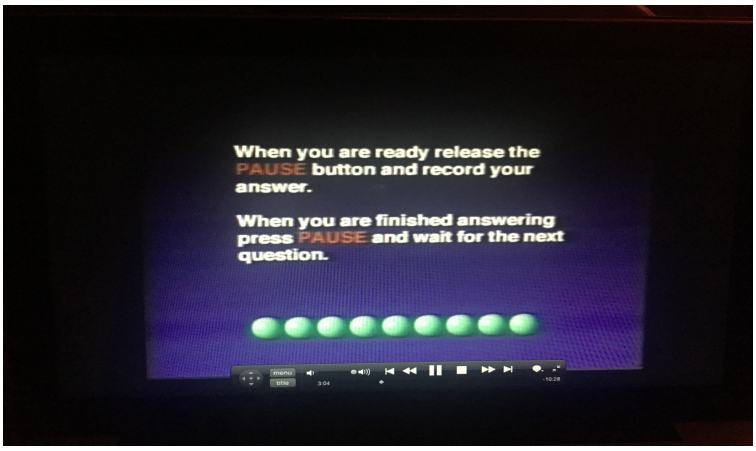

Figure 1. VOCI screenshot showing time remaining

The VOCI has a variety of questions that gradually progress from novice-level questions to more task-oriented questions at the intermediate, advanced, and superior levels. The questions have different functions, including description, comparison/contrast, role play, past tense narration, supporting an opinion, and hypothesizing. The questions take place in different, such as a restaurant, library, or coffee-shop. The following screenshot shows a restaurant situation, where one of the participants was eating his dinner. The other speaker turns to the camera and says “Ron is eating his dinner.” This is followed by a Novice-level question, “What do you eat for dinner?”

Figure 2. VOCI screenshot showing restaurant-themed question

Test questions take different forms. For example, there are situations where two people are talking, then one of them turns to the screen to ask the question, or both speakers ask parts of the question. For example, the following screen shot shows the two speakers sitting in a room reading newspapers, with the following dialogue:

Figure 3. VOCI screenshot showing past-tense narration question

A: It's really unbelievable.

B: Yes, that was a really terrific experience.

A: There are some experiences you just can't forget.

B: That's true. Have you ever had such an experience—an experience that you'll never forget?

A: It can be something positive or it can be something negative.

B: Tell us about it.

The function that this question is meant to elicit is past tense narration, an Advanced-level task. In other questions, no speakers appear on screen, but examinees can hear voices talking about a topic, and a picture that represents the topic appears on the screen. For example, with the image in Figure 4 below, test takers are asked: "Name at least five things that are represented in this picture." The purpose of this question is to elicit a list of items to be evaluated for vocabulary development.

Figure 4. VOCI screenshot showing vocabulary-themed question

Other questions are written on the screen, following a short dialogue. For example, the question depicted in Figure 5 began with a person sitting at an information

booth, saying “Hello, may I help you?” After that, information appears on the screen instructing the test taker to ask three questions, accompanied by a voice-over of those instructions.

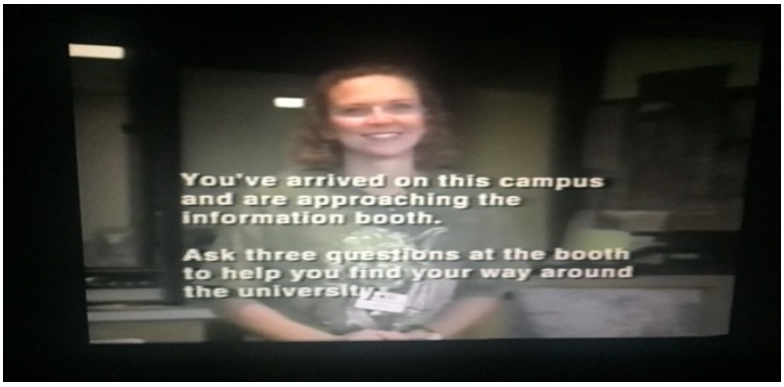

Figure 5. VOCI screenshot showing written questions on the screen Other test items illustrate a situation, and the question is presented as a voice-over. For example, Figure 6 shows a woman talking on the phone, accompanied by the following voice-over:

“Because of a last-minute problem, you missed a dinner engagement with a friend. You call to apologize, but your friend is not yet home, so you need to leave a message on the answering machine apologizing for the date and explaining why you were not there.”

Figure 6. VOCI screenshot showing apology-themed question with voice-over

TOEFL (Test of English as a Foreign Language)

TOEFL was introduced in 1963 by the National Council on the Testing of English as a Foreign Language. This test is for non-native speakers of English who need to demonstrate their proficiency, in order to be admitted to academic programs or considered for jobs in English-speaking countries. It is administered by the Educational Testing Service (ETS). There are two versions of the exam: an Internet-Based test (iBT) with four sections (reading, writing, speaking, listening) and a paper-based test (PBT) with two sections (reading, listening). Only the speaking section (of the iBT) will be described in this literature review.

The speaking section takes 17 minutes, with each section having a specific time, and is administered through the computer (ETS, 2019). Test takers are given four tasks, where they are asked to express their opinions regarding certain issues, and answer questions based on reading and listening tasks. Test takers' responses are recorded on the computer, then sent to ETS for rating.

What is an Utterance?

When it comes to measure oral proficiency, we use the utterance as the unit of measurement. Before measuring oral proficiency, we need to define which unit is being measured. Luoma (2004) emphasized that oral data has many sub-clausal units, especially in unplanned speech. Parsing oral data into units is considered very challenging, “as speakers hesitate, repeat, abandon topics, and reformulate their speech” (Vercellotti, 2012, p. 5). According to Crookes (1990), parsing oral data into units using linguistic features is preferred over parsing the data based on word counts. For example, using propositions, idea units, or c-units makes the coding process logical and consistent.

Although T-units have been used for both written and oral data (Foster, Tonkyn, & Wigglesworth 2000; Halleck, 1995; Norris & Ortega, 2009), Foster et al. (2000) claimed that many researchers modified T-units in order to use them in oral data. For this reason, Foster et al. (2000) introduced a new measure for oral data, which is called the Analysis of Speech unit (AS unit or ASU). An ASU (Analysis of Speech unit) is "a single speaker’s utterance consisting of an independent clause or sub-clausal unit, together with any subordinate clause(s) associated with it" (Foster et al. 2000, p. 365). Foster et al.’s AS unit is not only characterized by its syntactic criterion, but also aided with intonation and pause information.

The complexity of learner language can be seen in different ways. In written language, one looks at the complexity of sentences, or sometimes T-units; in oral language, one looks at the complexity of ASUs.

One way to analyze oral proficiency is through the general and specific measures of complexity, accuracy, and fluency. Those measures will be discussed in detail in the upcoming sections.

CAF Framework

Applied linguistics researchers have developed fine-grained measures that show the elements of learners' interlanguage in terms of complexity, accuracy, fluency (CAF) of production. Housen and Kuiken (2009, p. 461) described CAF measures as "performance descriptors for the oral and written assessment of language learners as well as indicators of learners' proficiency underlying their performance; they have also been used for measuring progress in language learning." Those measures have been used in a number of studies on language testing and language acquisition. CAF measures are dimensions and dependent variables that can be used to assess language performance in some language skills, mainly speaking and writing. One of the challenges in CAF measures is the fact that learner's performance according to those measures is very individualistic and may differ greatly from the group average (Larsen-Freeman, 2006).

The CAF framework is typically used in conjunction with Skehan's (1998) Trade-off Hypothesis and Robinson's (1995, 2001a) Cognition Hypothesis. The former states that learners assign varying amounts of mental resources to complexity, accuracy, and fluency during communicative tasks, and that increasing resources in one area results in a decrease to remaining areas. The latter provides a framework for conceptualizing elements of task complexity and argues for a careful, complexity-oriented sequencing of pedagogical tasks. Researchers rely on direct measures of CAF that take the form of

"ratio, frequencies, or formulas" (Norris & Ortega, 2009, p. 1). Based on Skehan (2009, p.510), successful performance in oral proficiency is dependent on 1) more advanced language, leading to complexity; 2) a concern to avoid error, leading to higher accuracy if this is achieved; and 3) the capacity to produce speech at normal rate and without interruption, resulting in greater fluency.

Complexity

Complexity in L2 production refers to the "size, elaborateness, richness and diversity of the learner's linguistic L2 system" (Hausen & Kuiken, 2009, p. 464). It is achieved through "expanding or restructured second language knowledge" (Wolfe-Quintero, Inagaki, & Kim, 1998, p. 4). It can be viewed as "the use of sophisticated forms (e.g., past passive modals), complex constructions (e.g., subordination), and various other late-learned production units" (Purpura, 2013, p. 119). On the syntactic level, complexity refers to the amount of subordination or other clausal measures, or mean length of unit of production (sentence or T-unit). Complexity is considered the most challenging construct in the CAF framework because it can be applied to lexical, interactional, propositional, and grammatical aspects (Ellis & Barkhuizen, 2005).

Raish (2017) claimed that learners' production becomes more complex when they are more proficient in the language. According to Gaies (1980), production is considered syntactically complex if it contains longer T-units.

Syntactic (grammatical) complexity.

Syntactic complexity has been defined differently by researchers. For example, Kuiken, Vedder, and Gilabert (2010) viewed production as syntactically complex when it contains a large number of clauses per T-unit. Several researchers used different measures for syntactic complexity. For example, Ellis and Yuan (2005) and Robinson (2007) used the raw tallies of specific verbal morphology (e.g. passive voice, tensed forms), syntactic structures (e.g. comparatives), or even classes of verbs (e.g. modals, or conditionals). Norris and Ortega (2009) claimed that since language can be elaborated at three different syntactic levels, three grammatical complexity measures must be used, which are global complexity (words per AS unit), phrasal complexity (words per clause), and complexity by subordination (clauses per AS unit).

Lexical complexity.

There are various methods to calculate lexical complexity. One method is determining the type-token ratio (TTR), which is the number of word types divided by all word tokens. Some researchers count the number of different word families, or the ratio of functional words to lexical words (Ellis & Barkhuizen, 2005). Another more complex method is Guiraud adjustment of TTR (e.g., Kuiken et al., 2010), in which the square root of the tokens is substituted for the number of tokens. Another adjustment for text length effects, the mean segmental TTR (MSTTR), determines the mean TTR of 100- word, or 50-word or 10-word segments of the text.

In a more detailed and precise description of complexity, Bulté and Housen (2012) illustrated the taxonomies of L2 complexity in the following diagram:

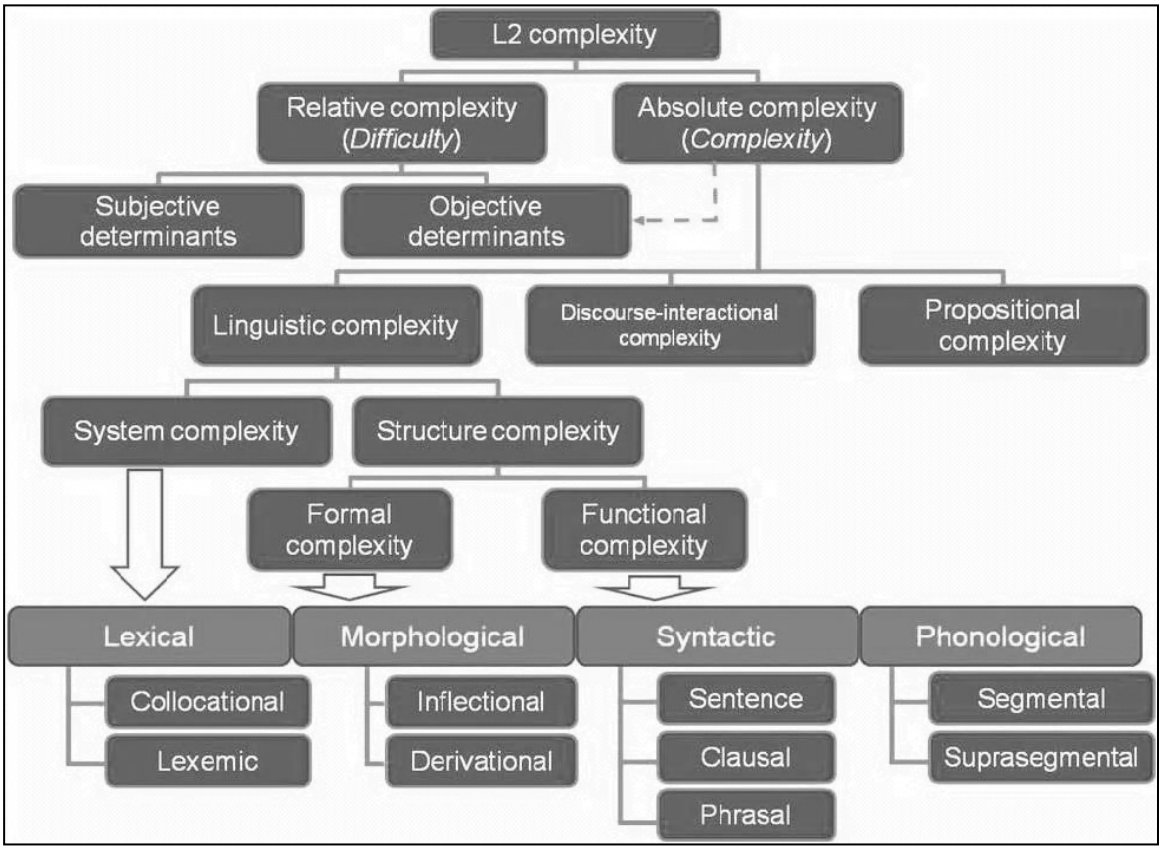

Figure 7. Taxonomic model of L2 complexity (Bulte & Housen, 2012, P. 44). Printed with permission from both authors through email.

Syntactic complexity is generally measured by length and subordination. The length-based measure uses the ratio of frequency of words to the total number of the chosen syntactic unit, most commonly the T-unit, which is defined as "an independent clause and all of its dependent clauses" (Iwashita, 2006, p.157). Subordination can be measured by "counting all clauses and dividing them over the chosen production unit" (Norris & Ortega, 2009, p. 558). Syntactic complexity also has more precise measures.

For example, Révész, Ekiert, and Torgersen (2016) propose measuring the frequency of certain syntactic forms, such as tense, aspect, verb patterns, or clause types.

In this study, I used two measures of syntactic complexity (number of Analysis of Speech Units (ASUs), and mean length of ASU), and three measures of lexical complexity (token counts, type counts, and type-token ratio). Those measures are the most widely used measures of complexity (Purpura, 2013).

Accuracy

Within the CAF framework, accuracy generally refers to the production of error-free discourse. It is the one measure about which researchers generally agree, in terms of its definition and operation (Housen, Kuiken, & Vedder, 2012; Pallotti, 2009). While the concept of error is very broad, researchers need to define what they mean by errors before doing any analysis. For example, Housen et al. (2012, p. 4) described errors as "deviations" from norms. Accuracy also includes grammatical accuracy (conformity of grammatical rules) and lexical accuracy (appropriateness of lexical items in particular context and for particular purpose).

One common way of measuring accuracy is dividing the total number of errors by the total number of words, especially for short speech segments. However, for long speech segments, the number of errors per, for example, 100 words, can be calculated (Mirshahidi, 2017). Another way of measuring accuracy is the proportion of error-free units (Tonkyn, 2012), which Foster and Skehan (1999) claimed is a reliable and sensitive measure of accuracy. Foster and Wigglesworth (2016) stated that accuracy can also be examined by looking at specific types of errors, such as tense-aspect errors, errors in

subject-verb agreement, or incorrect verb patterns. Accuracy has both general and specific measures (Vercellotti, 2012), which will be discussed in the following sub-sections.

Specific measures of accuracy.

Ellis and Barkhuizen (2005) define specific measures as "learner's suppliance of a specific form in obligatory contexts, which is best suited for focused tasks" (p. 151). Researchers can decide which forms to measure based on proficiency level. Robinson and Gilabert (2007) claim that specific measures should support general measures in order to capture the impact of resource-directing tasks. For example, when focusing on time and motion, Robinson, Cadierno, and Shirai (2009) used two accuracy measures centered on motion verbs, verb particles, and verb satellites.

General measures of accuracy.

Ellis and Barkhuizen (2005) recommend a general measure of accuracy, such as percentage of error-free clauses or number of errors per 100 words. Crookes (1990) believes that one advantage of using errors per 100 words is that the measure is not complicated by the difficulty of coding a clause, t-unit, or AS (speech) unit. However, this method is disadvantageous in the sense that 100-word segments do not have psycholinguistic reality while segments that are based on idea units, clauses, and AS units do (Crookes, 1990). General measures of accuracy are suitable for loosely structured data, such as data collected from informal conversations, where participants avoid using formal structures that include clauses or T-units, and for collecting data from speakers of

different languages (Vercellotti, 2012). Several researchers used both general and specific measures of accuracy. For example, Ahmadian and Tavakoli (2011) found that students in the careful online planning conditions had higher accuracy, which was measured generally by error-free clauses and specifically by verb forms. Also, Yuan and Ellis (2003) found the online planning group had significantly more accurate narratives than the no planning group, which was measured by error-free clauses and by correct verb forms. Other researchers have claimed that specific measures of accuracy can be very challenging and sometimes is based on a subjective coding. For example, Ellis and Barkhuizen (2005) claim that an ungrammatical utterance can be corrected in more than one way because the coder does not know the speaker's intended meaning. This makes global measures of accuracy a more realistic and sensitive measure (Skehan & Foster, 1999).

According to Larsen-Freeman (2009) the best global measures of accuracy are "the number of error-free T-units, error-free T-units per T-unit, and errors per 100-word" (p. 580). In this study, I measured accuracy in terms of the percentage of error-free ASU.

Fluency

Fluency has been defined by De Jong, Groenhout, Schoonen, and Hulstijn (2015) as "speedy and smooth delivery of speech without (filled) pauses, repetitions, and repairs" (p. 224). They further categorized fluency into three major types of cognitive, perceived, and utterance fluency. Utterance fluency seems to be the most tangible concept of the three. Measures of utterance fluency will be discussed in the upcoming section.

Measures of utterance fluency.

Utterance fluency can be measured by counting the number of filled pauses and their durations, corrections, and repairs. This leads us to the three subcomponents of utterance fluency, which are speed (rate of speech), breakdown (silence and pausing), and repair fluency (hesitations and reformulations) (De Jong et al., 2015; Segalowitz, 2010). Some researchers have distinguished the speed and the flow fluencies, such as Skehan (2014), who claimed that dis-fluencies that interrupt flow should be distinguished from dis-fluencies that influence speed (Tavakoli, Campbell, & McCormack, 2016). Skehan's fluency framework illustrates two major types of fluency - speed and flow. The former can be measured in terms of speech rate, and the latter can be measured using pausing and reformation (Skehan, 2014).

Tavakoli (2016) and Witton-Davies (2014) provided a detailed description of measuring utterance fluency in terms of pauses and speed. While pauses include length of pause, frequency of pause, and location of pause in the clause, speed refers to speech rate and articulation rate, which could also include phonation time (i.e. speaking time minus pauses), mean length of run (i.e. mean number of syllables between pauses), and repair measures (e.g. number of hesitations, reformulations, etc.). In this study, I used two measures of breakdown fluency (silent pauses and filled pauses). Following previous scholars (Freed, 2000; Freed et al., 2004; O'Brien, Segalowitz, Freed, & Collentine, 2007; Tavakoli & Foster, 2008; Tavakoli & Skehan, 2005), I used 400-milliseconds as a cutoff. Any pause less than 400-milliseconds was not identified as a silent pause, because native and non-native speakers commonly make short pauses in their speech.

I argue here that breakdown fluency is a more valid fluency measure than speed fluency or repair fluency measures. I believe that these two latter measures might not be related to oral proficiency but could be related to personal style in speaking. For example, some people speak quickly by nature, even in their native language, while others do not. Similarly, speakers can sometimes have issues of repair fluency (hesitation, self-repetition). In this dissertation, I used two measures of breakdown fluency, frequency of silent and filled pauses. Vercellotti (2012) stated "talking quickly is not the point, as speaking teachers will stress. A non-fluent speaker is not specifically identified by slow speech but by a breakdown of fluency" (p. 22).

Summary of the CAF Measures

This section provides a broad description of the three measures as reviewed by Ellis and Barkhuizen (2005). Tables 1, 2, and 3 provide definitions of the measures and the previous studies that used them.

Table 1.

Measures of Complexity (Ellis and Barkhuizen, 2005, pp. 153-154)

	Measure	Definition	Study
Interactional	Number of turns	The proportion of the total number of turns in the interaction for each speaker.	Duff (1986)
	Mean turn length	The total number of words divided by total number of turns.	
Propositional	Number of idea units encoded	The total number of major and minor idea units.	Zaki and Ellis (1999)
Functional	Frequency of some specific language function	The total number of times a specific language function is performed by a learner.	Brown (1991)

Measure		Definition	Study
Grammatical	Amount of subordination	The total number of separate clauses divided by the total number of c- (or AS) units.	Foster and Skehan (1996)
	Use of specific linguistic features.	frequency of specific linguistic features.	Yuan and Ellis (2003)
	Mean number of verb argument	The total number of verb arguments divided by the total number of finite verbs.	Bygate (1999)
Lexical	Type-token ratio	The total number of types divided by the tokens.	Robinson (1995)

Table 2.

Measures of Accuracy (Ellis and Barkuizen, 2005, p.150)

Measure	Definition	Study
Number of self-correction	The number of self-correction as a percentage of the total number of errors committed.	Wigglesworth (1997)
Percentage of error-free clauses	The number of error-free clauses, divided by the total number of independent clauses, sub-clausal units and subordinate clauses, multiplied by 100.	Foster and Skehan (1996)
Errors per 100 words	The number of errors, divided by the total number of words produced, divided by 100.	Mehnert (1998)
Percentage of target-like verbal morphology	The number of correct finite verb phrases, divided by the total number of verb phrases, multiplied by 100.	Wigglesworth (1997)
Percentage of target-like use of plurals	The number of correctly used plurals, divided by the number of obligatory occasions for plurals, multiplied by 100.	Crookes (1989)

Measure	Definition	Study
Target-like use of vocabulary	The number of lexical errors, divided by the total number of words in the text (excluding dysfluencies).	Skehan and Foster (1997)

Table 3.

Measures of Fluency (Ellis and Barkhuizen, 2005, p. 157)

Measure		Definition	Study
Temporal	Speech/writing rate	Pruned syllables (i.e. excluding dysfluencies) is counted and divided by the total number of seconds/minutes the text or utterance took to produce.	Ellis (1990b)
	Number of pause	The total number of filled and unfilled pauses for each speaker.	Robinson, Ting, and Urwin (1995)
	Pause length	This can be measured as either the total length of pauses beyond some threshold (e.g. 1 second), or as a mean length of all pauses beyond the threshold.	Skehan and Foster (1999)
	Length of run	This is the mean number of syllables between two pauses of a pre-determined length (e.g. 1 second). This measure discounts dysfluencies.	Wiese (1984)
Hesitation phenomena	False starts	Utterances/sentences that are not complete (i.e. constitute fragments). They may or may not be followed by reformulations.	Skehan and Foster (1999)
	Repetitions	Words, phrases, or clauses that are repeated without any modification whatsoever.	
	Reformulations	Phrases or clauses that are repeated with some modification	
	Replacement	replaced lexical items	

Effects of Task Types on CAF Measures and Oral Proficiency Performance

Although some studies have examined the influence of task types on oral production, Ellis (2009) claimed this area is still under-researched. Some researchers have found that that for tasks requiring simple and familiar information, learners' production is more fluent and accurate, but less complex (Foster & Skehan, 1996; Mehnert, 1998). Foster and Skehan (1996) examined the effects of three different tasks (personal information exchange, narrative, and decision-making) on fluency, complexity, and accuracy. Their findings indicate that planning had more influence on narrative and decision-making tasks than on the personal information exchange task. In contrast, other studies have found that for tasks requiring higher communication pressure or textual demand, learners increase their accuracy in certain grammatical forms (Tarone, 1985; Tarone & Parrish, 1988). In addition, Bygate (1999) compared oral proficiency performance in narrative and argumentation tasks. For narratives, learners produced more verb arguments and relative clauses, while for argumentation, they produced more verb groups, individual verb forms, and nominal clauses. In another study, Hu (2018) examined the effect of task type on oral second language production. She used two tasks, relating to a map and a picture-story. Her findings revealed that picture-story production had higher syntactic complexity and lexical diversity, while map task production had simpler, shorter, and less varied language.

Skehan and Foster (1999) investigated the effects of task structure and processing load on learner performance on a narrative retelling task. They found that structured tasks generated more fluent language, and complexity of language was affected by the processing load. Moreover, Lumley and O'Sullivan (2005) examined the effect of different variables (gender, task type, and topics) on 894 students from two Hong Kong universities on a tape-based test. They found that some tasks and topics might be more appealing to males than females, which can affect their oral proficiency performance. In addition, Huei-Chun (2007) examined the effect of task type on the performance of 30 Taiwanese students, using a semi-direct speaking test and a questionnaire. Her findings indicate that there are significant main effects for complexity.

Studies That Compared Oral Proficiency using Direct and Semi-direct Testing Instruments

Before reviewing the studies that examined direct and semi-direct tests of oral proficiency, definitions of those types of tests should be illustrated. In direct tests, a live examiner conducts a face-to-face interview with the examinee. In semi-direct tests, examinees receive prompts from multimedia sources, and their responses are recorded and later assessed by trained raters (Alderson & Banerjee, 2002; Jeong et al., 2011; O'Loughlin, 2001; Qian, 2009; Shohamy, 1994). Qian (2009) has noted that the semi-direct setting is more practical in most settings, as on-site, expert examiners are not required to conduct the test.

The literature is replete with studies that compare oral proficiency in direct and semi-direct testing modes. These studies can be broadly classified into three groups:

studies examining test takers' performance, studies examining test takers' perceptions, attitudes, or preferences, and studies examining test validity and other test features.

Zhou (2015) compared two monologic tasks (narrative and opinion) delivered by a computer and a face to face interview. Zhou's findings showed evidence for the validity of computer-delivered monologic tasks, which means that there was not any difference in test scores in both testing modes. Zhou suggested that

> "the results on computer-delivered monologic tasks could be used to infer scores on face-to-face monologic tasks. Moreover, the same underlying factor structures measured by monologic tasks in the two modes suggest that scores on computer-delivered monologic tasks could be interpreted similarly to those scores on face-to-face monologic tasks." (p. 14)

Zhou encouraged future studies to use different tasks, other than opinion and narrative. In the computerized testing mode, Zhou allowed the participants to have preparation time before recording their answers, which was addressed as one of the limitations of this study, as this led participants to rehearse their performance. The present study will use other types of tasks, including comparison, description, and role play, with no preparation time given to examinees.

Another study by Brown, Cox, and Thompson (2017) compared performance on the in-person OPI and computer-mediated OPIc, focusing on lexical diversity, lexical density, and discursive features. Specifically, they compared the performances of examinees who scored Advanced Mid level on both tests with examinees who scored Advanced Mid on the OPIc and Advanced Low on the OPI. When comparing

performance on a tense narration task in both tests, they found that for temporal fluency, participants' speech rate was significantly higher on the OPI. Duration of silent pauses was also longer on the OPI, although the total number of silent pauses was not significantly different. For repair fluency, examinees had more verbatim repetitions and false starts on the OPI, but the number of filled pauses was not significantly different. As for lexical diversity and density, the OPIc elicited greater lexical diversity and density from the testees than did the OPI.

Jeong (2003) examined the multimedia-enhanced oral proficiency interview (d-VOCI) based on students' electronic literacy questionnaire and their scores in OPI. Jeong used the Korean version of VOCI, not the English version. She had 144 Korean college students. The purpose of her study is to test the possibility of using d-VOCI in a Korean college classroom as the multimedia-enhanced oral proficiency test in Korea as a new method for language testing. One of the questions that Jeong addressed is comparing oral proficiency scores in d-VOCI and face-to-face interviews. Her findings show that the mean scores of OPI and d-VOCI are different (M=16.40, M=4.12, respectively) (Jeong, 2003, p.71). Also, the statistical results showed a weak relationship between these two tests (.30). Jeong considered this correlation weak because it indicated that knowing the results of students' face-to-face interview does not contribute in predicting students' d-VOCI scores (9%). She also found that while there was a low inter-rater reliability between the face-to-face interview raters (.64), the inter-rater reliability of d-VOCI raters was high (.90) (Jeong, 2003, p. 102). One of the reasons for these differences between face-to-face interviews and d-VOCI in terms of inter-rater reliability is the possibility of using different rubrics and different formats for scoring. Also, Jeong believes that lack of

evaluators' training could be one of the factors that affect the low inter-rater reliability in face-to-face interviews. She also thinks that whereas in face-to-face interviews, the raters usually take notes during the interview which they use later for rating, in the d-VOCI, raters record the speech samples using audio files and CDs, which can be accessed at any time for a more valid evaluation than OPI. Jeong did not compare the CAF measures in both conventional interview and d-VOCI, which leaves a gap that this current study aims to fill. Giles (2016) encouraged researchers to compare fluency and complexity in three different testing modes: monologue, student-to-student dialogue, and examiner-student dialogue. This current study contributes to the literature of oral proficiency assessment by comparing complexity, accuracy, and fluency in two testing modes, a monologic test (VOCI, and participant-computer communication) and dialogic test (OPI, participant- and examiner communication).

Other studies have examined the effect of testing mode on oral proficiency. For example, Chapelle and Douglas (2006) highlighted the limitations of technology-based proficiency tests and claimed that they produce results that do not match those generated by other assessments. Chapelle (2003) has urged researchers to compare computer-based tests with conventional tests, and Alderson (2004) has called for more research on the influence of computer-based testing.

Very few studies that examined test takers' performance in terms of the psychometric properties of computer-delivered speaking tests and face-to-face tests. For example, Malabonga, Kenyon, and Carpenter (2005) have examined the effectiveness of technical aspects of the Computerized Oral Proficiency Instrument. Also, Swain, Huang, Barkaoui, Brooks, and Lapkin (2009) investigated examinees' strategic behaviors on the

Speaking section of the Internet-based Test of English as a Foreign Language (TOEFL iBT). Other researchers including Brooks and Swain (2014) have compared test takers' performance on the speaking section of the TOEFL iBT with their actual academic performance during speaking activities in classes and outside classes. Their participants reported that they were more engaged during the speaking activities that were conducted face-to-face because they were not thinking about any grammar or vocabulary usage. Among those studies that compared these two types of testing modes is Zhou (2008) that focused on test takers' speaking samples. In his study, he found that while examinees utilized more repetitive words during the interviewer-delivered monologic tasks, they used more filled pauses during the computer-delivered monologic tasks. Another study was conducted by Jeong, Hashizume, Sugiura, Sassa, Yokoyama, Shiozaki, and Kawashima (2011) who concluded that direct interviews might elicit a more varied communicative ability than semi-direct interviews.

Studies comparing OPI and SOPI have found that examinees' scores in these two testing modes are equivalent. For instance, Shohamy (1994) examined 10 participants and found no difference in mean scores between the Hebrew OPI and SOPI. Similarly, Kenyon and Tschirner (2000) found no differences between scores on the German OPI and SOPI for his 20 participants. The ratings in these studies were based on the ACTFL Guidelines, a holistic rating scale. Therefore, according to Zhou (2015) "investigations of the differences between modes using analytic scales have not been conducted" (p. 3). Zhou (2015) further pointed out that the sample size in both studies was small. In addition, Shohamy (2004) did not make it clear whether she compared the same group of subjects, and Kenyon and Tschirner (2000) did not use a counterbalanced design, which

makes practice or fatigue possible affective variables. Zhou (2015) also believed that since the OPI and the SOPI differ in terms of task type and content, it has not yet been determined if comparisons between these two tests are valid.

Previous studies investigated the validity of different types of direct and semi-direct tests. For example, studies examining the concurrent validity of test scores (e.g., Stansfield, 1991; Stansfield & Kenyon, 1992) found that the two testing modes are equivalent, with no difference in concurrent validity. Other studies analyzed the effect of testing modes on pragmatic or discourse features (e.g., Luoma, 1997; O'Loughlin, 1997, 2001; Shohamy, 1994), and found key differences between the two testing modes, including test takers' preferences and performances. For example, Shohamy (1994) found significant differences in communicative strategies, with participants using more grammatical self-corrections and paraphrasing in the SOPI, and more L1 code switching in the OPI. Luoma (1997) found the OPI and SOPI testing modes to be very similar in terms of examinee scores and linguistic forms, but significantly different in the usage of parts of speech and individual words. Studies on test takers' attitudes and perceptions towards these two testing modes have produced mixed results. They revealed that although most test takers prefer face-to-face (direct) testing (McNamara, 1987; Qian, 2009; Shohamy, Donitsa-Schmidt, & Waizer (as cited in O'Loughlin, 2001, p. 29); Stansfield et al.,1990), some prefer semi-direct tests (Brown, 1993; James, 1988). However, studies comparing examinees' perceptions towards these testing modes is still described as "limited" (Suryaningsih, 2014, p. 5).

Other studies have examined test takers' attitudes towards IELTS. Ata (2015) investigated the attitudes and perceptions of Chinese, Arab, and Indian students in Australia, using a questionnaire. He found that Chinese students were more likely than Arab and Indian students to agree that "lack of confidence and stress are major factors for them" (p. 496) during the exam. Rasti (2009) examined the perceptions of Iranian test-takers towards IELTS using an attitude questionnaire (60 participants) and semi-structured interview (12 participants). His findings revealed that, overall, 80% of participants had a positive attitude towards IELTS. Half of the candidates found it relaxing to take the exam and felt self-confident when being interviewed. Winke and Lim (2014) examined how test anxiety affected performance on the IELTS listening section. They determined that test familiarity could cause anxiety and poor test management skills for test takers.

Few studies have compared IELTS to TOEFL in terms of test takers' attitudes. Gardiner and Howlett (2016) examined students' perceptions towards IELTS, TOEFL iBT, and two other university gatekeeping tests. Test takers reported having more anxiety in IELTS, because they had to generate ideas in the presence of an interviewer. They also cited noise disturbances as the only external factor affecting their performance on the TOEFL iBT. Suryaningsih (2014) compared perceptions towards TOEFL and IELTS by conducting semi-structured interviews with six graduate students. All participants viewed the IELTS positively in terms of testing experience, perceptions of time, perceptions of task, and effects of the test, but viewed the TOEFL negatively with regard to these themes. All participants had negative perceptions toward both exams with regard to test topic and test score.

Qian (2009) compared perceptions toward a direct test (IELTS) and semi-direct test of oral proficiency (speaking component of the Graduating Students' Language Proficiency Assessment–English). Her 243 participants were final-year students at a Hong Kong university. She found that participants who strongly favored direct testing outnumbered participants who strongly favored semi-direct testing. For participants who disliked semi-direct testing, lack of interaction with the examiner was the main reason. Qian (2009) summarized the findings of some of the previous studies that examined the test takers' attitude towards direct, semi-direct and indirect testing modes (See Table 4).

Table 4.

Test Takers' Reactions toward Testing Modes (Percentage Based) (Qian, 2009, p.117)

Study	Test	In favor of semi-direct test	Neutral	In favor of direct test
Brown, 1993	Occupational Foreign Language Test (Japanese)	57%	18%	25%
James, 1988	Test in English forEducational Purposes	Many	N/A	N/A
McNamara, 1987	Occupational English Test	30%	18%	52%
Shohamy et al., 1993	Hebrew OPI and SOPI	4%	—	96%
Stansfield et al., 1990	Portuguese Speaking test	7%	7%	86%

The present study responds to the limitations and suggestions found in previous studies. For example, all of the previous studies used different types of direct and semi-direct tests of oral proficiency, but none examined the English version of the VOCI or compared it with other tests. Only one study (Jeong, 2003) compared the Korean version of the VOCI with the OPI in terms of oral proficiency. In addition, Jeong et al. (2011) encouraged future researchers to use different types of oral proficiency tests and to recruit participants of different proficiency levels, as their study had only used intermediate level participants. The present study responds to these suggestions by using the OPI and VOCI and including advanced L2 speakers. Alderson (1988) recommended using test takers' feedback from as many sources as possible. The present study therefore uses Saudi test takers, a population that was not used extensively in the previous literature, to examine their perceptions towards direct (OPI) and semi-direct (VOCI) tests.

Hill (1998) found that participants who were native speakers of Asian languages were more nervous during a face-to-face interview than a tape-based test. She hypothesized that these participants may have been "less familiar with communicative language learning techniques and therefore more comfortable with the relatively predictable and structured format of the language laboratory than they are in a face-to-face interview" (p.218). It is worth investigating this issue using a different population with a different language background.

The present study is significant because these two tests (VOCI & OPI) use the same functions and guidelines, as the VOCI was constructed based on ACTFL oral proficiency guidelines (Halleck, personal communication). Similarly, the OPI tester is a certified ACTFL rater who is trained to use the ACTFL guidelines during the OPI.

However, the two tests are different in terms of the mode of delivery, where the OPI is a direct face-to-face interview and the VOCI is a semi-direct test delivered through the computer. It would be interesting to find out if the mode of delivery would make a difference for test takers in terms of their preferences and perceptions toward these two tests. This study further compares test takers' preference with their testing performance, examining issues of complexity, accuracy, and fluency. Qian (2009) found that his participants believe that they performed better in the testing mode they felt more comfortable with.

It would be interesting to examine the previous assumption made by Qian (2009) and determine whether Saudi ESL participants perform differently in the testing mode they prefer.

The current study will add to the very limited literature that has examined VOCI in assessing oral proficiency. It will be only the second contribution using an English VOCI. This can draw interest from researchers and language educators in exploring more about this testing instrument. The study will also provide useful information to test developers who wish to know why test takers prefer one test over another, and how testing modes might affect preferences, perceptions, and performance in terms of CAF measures.

CHAPTER III

METHODOLOGY

Overview of the Chapter

This chapter presents the methodologies used to conduct the study. It begins with a description of the participants, then provides details of the instruments used in the study (VOCI, OPI, online background survey, and Arabic interviews). It concludes with a description of the study procedures.

Participants

A group of Arabic-English bilinguals were recruited using convenience sampling. Thirteen male engineering students from Saudi Arabia participated in this study. However, based on the ACTFL proficiency guidelines, only those at an advanced level of oral proficiency were chosen for data analysis. Participants were all third- or fourth-year undergraduate students majoring in different fields in the College of Engineering, at a university in the south-central United States. All participants had taken IELTS and TOEFL in order to be admitted to their academic programs, and all were considered advanced ESL students based on proficiency descriptions in the ACTFL guidelines. Their mean age was M= 23.78. The following table shows background information for all nine participants.

Table 5.

Participants' Background Information

Name	**Age**	**Major**	**Native language**	**Native country**	**Length of stay in USA**	**Years of studying English**
Talal	25	Chemical engineering	Arabic	Saudi Arabia	5	5
Alabad	30	Chemical engineering	Arabic	Saudi Arabia	3	15
Alali	22	Mechanical engineering	Arabic	Saudi Arabia	5	6
Ahmad	23	Civil engineering	Arabic	Saudi Arabia	2	4
Nasser	24	Chemical engineering	Arabic	Saudi Arabia	5	7
Odai	22	Mechanical Engineering	Arabic	Saudi Arabia	5	7
Salman	23	Mechanical Engineering	Arabic	Saudi Arabia	5	8
Aziz	23	Industrial engineering	Arabic	Saudi Arabia	5	11
Mohamed	22	Mechanical Engineering	Arabic	Saudi Arabia	3	2

Instruments

Four instruments were used to conduct this study, Video Oral Communication Instrument (VOCI), Oral Proficiency Interview (OPI), online survey, and Arabic interviews. Each is discussed in detail in the sections below.

VOCI

The English version of the Video Oral Communication Instrument (VOCI) is an oral proficiency test that was developed in 1995 by Halleck and Young, for the San Diego State University Language Acquisition Resource Center. VOCI incorporates both visual and audio input, presented through an audiovisual tape or computer file. VOCI uses technology to collect speaking samples on carefully designed tasks, which are rated based on the ACTFL scale (Kenyon, 1998). It is offered in seven languages, including English. The test consists of 35 questions that assess four levels: novice, intermediate, advanced and superior. The first three questions were mainly for acquainting test takers with the test and ensuring that sound and pictures are clear. The test provides different situations or scenarios, which are followed by a question for test takers.

VOCI can be administered individually or in groups and can be timed or untimed. This study used a timed (46-minute), individual version that was digitally recorded for data collection. Participants took the full exam on a MacBook Pro computer, while alone in a quiet room in the university library. In the timed version of the exam, there are green bouncing balls that decrease in number to represent the remaining time for the given task. The test taker has to finish speaking before the balls disappear. The complete list of questions can be found in Appendix A.

OPI

Oral proficiency interviews (OPIs) were also conducted individually with a certified ACTFL tester, who also appears in the VOCI as one of the discussants. The interview lasted approximately 30 minutes, and took place in the interviewer's office in the university's English department. The test was recorded on a digital recorder. The interview includes questions that ranged from novice level to superior. Below are some of the questions that were asked during the OPI:

1) Where are you from? (Novice)
2) Choose one place that you went to that you enjoy and tell me about it. (Intermediate)
3) Can you compare your city in Saudi Arabia with a city in the US? (Advanced)
4) Let's do a role play. OK, here's the situation imagine that you've been living here for a number of years and you have worked hard to promote multicultural awareness and understanding of people from different places and you can you get an award you win the student of the Year Award for promoting multicultural awareness, and OSU gives a luncheon in your honor. And you have to accept this award. So I'll introduce you and clap and you can make your very brief acceptance speech. OK. Ladies and gentlemen I'd like you to welcome this year's winner of the multicultural award. Give him a warm welcome. (Superior)

Online Background Survey

The researcher designed a 10-question background survey that was conducted through SurveyMonkey. The topics were as follows:

- Pseudonym (Q1)
- Age (Q2)
- Class level (Q3)
- Academic major (Q4)
- Native language (Q5)
- Length of studying English in years (Q6)
- Length of living in the United States (Q7)
- Type of standardized tests taken before (Q8)
- Type of test you have taken in which you got a higher score in the speaking part (Q9)
- How do the tests you have taken; the face to face interview and the video, compare to your testing experience with other tests you have taken before? Which one or ones do you prefer? (Q10)

Arabic Interviews

The Arabic interview was the last instrument used. It consisted of only one question. However, I should mention here that a few months after data collection, participants were contacted again through personal communication to ask them some specific questions about their answers.

من خلال اختبارات مهارة التحدث اللي اختبرتها قبل كذا، هل تفضل الاختبار وجه ل وجه ام عن طريق الكمبيوتر؟ وماهي الأسباب

Among the other tests you have taken, which method of testing do you prefer, face to face or computerized testing, and why?

Procedures

Participants were recruited in the spring of 2018, using convenience sampling through friends. I met with recruited participants to explain the purpose of the study and what participation would entail, emphasizing that they would need to complete all four stages of the study (OPI, VOCI, background survey, Arabic interview). Participants

signed a consent form and were told that an iPad would be given to one participant, selected at random, who completed the entire study. Appointments were then scheduled for the OPI. After completing the OPI, participants scheduled an appointment to take the VOCI. Once they had taken both exams, participants were sent the background survey via email. Months later, they completed the Arabic interview.

Data analysis began with transcription of the VOCI and OPI, using Microsoft Word and InqScribe transcription software. After that, four types of questions were chosen for analysis (comparative, descriptive, role play, and past tense narration) from both OPI and VOCI. Those questions were chosen because ,based on the ACTFL guidelines, advanced level speakers should be able to do those functions (comparison, description, role play, and narration). The researcher used two criteria to determine which questions could be used for analysis. First, questions should have the same function, meaning they required participants to perform the same tasks (compare, describe, role play, narration). Second, they should be either about the same topic or at the same level of difficulty.

The VOCI/OPI coding process had several steps. First, all sentences were parsed into Analysis of Speech units (ASUs), which refers to any independent clauses with all of its dependent clauses (Foster et al., 2000), and the number and length of ASUs were calculated per 100 words. Next, the number of tokens and type of tokens were counted, and the Mean Segmental Type-Token Ratio (MSTTR) per 100 words was calculated. I did not delete filled pauses “ah” and “like” as they are part of the participants’ tokens. It is worth mentioning that all of the complexity measures were coded twice by the researcher, first manually in MS Word, then by using InqScribe. After that, another PhD

candidate in Linguistics coded the responses of five participants, and an inter-rater agreement of 97% was reached.

Responses were then analyzed for accuracy in terms of grammatical and lexical errors, based on the grammaticality of ASUs. To perform this analysis, the researcher read each ASU, underlined any words that might be erroneous, then decided which grammatical and lexical aspects are affected in the underlined words (e.g., tense, plurality, subject-verb agreement). After that, another rater performed the same procedures, with an inter-rater agreement of 92%. As lexical errors were rather limited, the researcher decided to examine only grammatical errors in the study. Grammatical errors refer to any deviation from standard English, such as tense and subject-verb agreement, misusing prepositions, and subject and verb deletion.

Measures of complexity and accuracy were coded twice. A graduate candidate in Linguistics transcribed five participants' VOCI responses and another three participants' OPIs responses, in terms of the number of ASUs and identifying the grammatical errors. Inter-coder reliability was calculated for the responses of five participants in all of the tasks they did and that was for coding the number and length of the ASUs, and the Error-free ASUs using two-way intraclass correlation coefficient ($p<.01$), which is shown in the following table:

Table 6.

Inter-coder Reliability

Task	Intercoder reliability	
	ASUs	**Error-Free ASUs**
Comparison	0.94	0.87
Description	0.95	0.91
Role play	0.97	0.94
Past tense narration	0.93	0.95

Fluency measures took more steps to evaluate than other measures. Digital audio files from the OPI and VOCI were converted into WAV files, through the website https://online-audio-converter.com, then opened in Praat, a software for phonetic analysis. Sound files for the questions selected for analysis were extracted as separate wav files, then opened again in a new Praat window. These extracted files were then annotated into a text-Grid that includes phonation rate, silent pauses, time spent talking, and total time (including pauses and sounding). From those measures, I have chosen silent pauses and pauses filled with “like” and “ah”. These two common gap fillers were used differently in the two testing modes, in those questions selected for analysis.

Several steps were followed in order to calculate the CAF measures. The following paragraphs will illustrate the steps taken in calculating each CAF measures.

Complexity was measured in terms of the Mean Length of the Analysis of Speech Units (ML-ASU), and the Mean Segmental Type Token Ratio (MS-TTR). As for the

ML-ASU, first, I segmented the responses into ASUs, which refers to all dependent clauses with all of their independent clauses. Then, I segmented those ASUs into 100 words. After that, I counted the number of ASUs per 100 words. Then, I counted the total number of tokens three times (manually, using word counts in word documents, and using text-inspector website). After that, I calculated the mean length of ASU by dividing the total number of tokens by the total number of ASUs per 100 words. As for MS-TTR, using the 100-word segments, I counted the type-token ratio for each segment. Then, the mean score for all TTRs was calculated.

Accuracy was measured using percentage of error-free ASUs. Using the 100-word segment, I underlined all grammatical errors. Then, I counted the number of ASUs that had grammatical errors. After that, I subtracted the number of ASUs with grammatical errors from the total number of ASUs, per 100 words, and that gave us the number of EF-ASU. Then, I divided the number of EF-ASUs by the total number of ASUs multiplied by 100.

Fluency was measured in terms of the number of silent and filled pauses. Using the 100-word segment, the silent pauses that are greater than 400ms, and that was measured using Praat as discussed before, were circled. Then, any instances of "ah" and "like" that function merely as gap fillers were put between square brackets. A reader who had not listen to the examples might think that some instances of "like" was not used as gap filler; however, I used my judgment as a listener to decide that it is a gap filler, as this depends on how the speaker says it.

An example of CAF measures analysis is shown in the following illustrations. The example is taken from the VOCI for the participant (Odai). The question was "Compare your hometown with a city that you have visited recently or know very well." Table 7 shows the symbols and notation that I used to analyze the responses in terms of CAF measures:

Table 7.

Symbols and Their Meanings

Symbols	**Meaning of the symbols**
Slash /	Boundaries of ASUs
Grammatical	Grammatical errors
[>400ms]	Silent pauses greater than 400ms
{ }	Filled pauses
…	Incomplete thought

One challenge for all CAF analysis was determining whether a silent pause was part of the previous ASU, or part of the new ASU. Throughout the data analysis, I placed all the silent pauses at the beginning of ASU or the beginning of the clause, based on the assumption that participants had finished talking about their last ASU and were pausing to think about the next one.

An example of CAF measures analysis is shown in the following illustrations. Example 1 is taken from the VOCI for the participant Odai. The question was "Compare your hometown with a city that you have visited recently or know very well."

Example 1 – Describe your hometown (Odai)

I am going to compare Stillwater to my city in Saudi Arabia/ my city is Jeddah/ there are many [>400ms] differences between the two places/ {ah} Jeddah is a big and modern city/ but Stillwater is just a small town/ {ah} in Jeddah there are many factories, big companies, restaurants, businesses, and a lot of fun places/ [>400ms] it is on the Red Sea/ so it is very humid/ Stillwater is just like a college town/ it does not have companies or big…../ it is not a business city/ {ah ah} I think the similarity between Stillwater and Jeddah is that both places [>400ms] has (grammatical) diversity/ but Jeddah is more {ah} diverse/ there is only one college in Stillwater/ but Jeddah has {like} at least 10 colleges/ some of the colleges are private/ and some are {ah} public/ Jeddah is very crowded/ and it has so much traffic/ Stillwater is not crowded/ you can go anywhere in less than 15 minutes/ [>400ms] also Jeddah is expensive/ I mean living there is expensive/ but Stillwater is way cheap (grammatical)/ the good thing about Stillwater is that it has clean air /so it is healthier to live here/ however, Jeddah is {ah} polluted because of the cars, engines, and factories/ {ah ah} I think the differences are more than the similarities

In this example, based on 100 words, the number of ASUs is 28, the ML-ASU is 6.97, and the MS-TTR is 0.59. MS-TTR was calculated by averaging the TTR for the first 100 words (0.56) and the second 100 words (0.63). The percentage of error-free ASUs is 93%, which was obtained through calculating the EF-ASU for the first 100 words (1 error in 12 ASUs) and the second 100 words (1 error in 16 ASUs). The

grammatical errors are related to subject-verb agreement, adverbs, and comparative forms. There were also 4 silent pauses and 10 filled pauses (9 instances of "ah" and one instances of "like").

……Stillwater and Jeddah is that both places [>400ms] has diversity

↓

……have…..

After analyzing the CAF measures in participants' OPIs and VOCIs, the researcher organized the online survey responses into a table, entered their responses to the background questions and the question about preferences toward OPI and VOCI.

In order to find the perceptions and preferences of the participants towards OPI and VOCI, I read through the transcribed data and decided on the themes. I chose the following categories that are directly related to the research questions: positive and negative perceptions towards the OPI, and positive and negative perceptions towards the VOCI**,** preferences towards OPI, preferences towards VOCI, reasons for the preference towards OPI, and reasons for preferences towards the VOCI. Following the coding procedures, I started reading through the data again, underlining all ASUs, writing whether the participants were talking about OPI or VOCI, deciding on the category that the sentence belongs to, and then giving each sentence a theme that explains what that sentence is about. After analyzing and coding all the transcripts, I found 22 themes. Upon further examination, I divided them into three major themes and nineteen sub-themes. Then, I looked at the themes and grouped the sub-themes that are related to each other

under the major theme. For example, the theme “interactions” has several subthemes that involve aspects of interaction, such as presence of a human being, tailored questions, examiner’s reaction, engagement/ involvement, and ambiguity/clarification/explanation. After that, I looked at the major themes and made sure that they are mutually exclusive.

Finally, statistical analysis was performed. Descriptive statistics were used to calculate the mean, standard deviation, and median. Then inferential statistics were conducted, using the Wilcoxon signed-ranked test and the Kruskal Wallis test. Those tests were chosen because they are non-parametric tests, where the Wilcoxon was used to examine the differences of CAF measures in the OPI and VOCI and the Kruskal Wallis was used to explore the impact of task type on CAF measures.

CHAPTER IV

RESULTS AND DISCUSSION

This section presents the findings of this study and discusses how they relate to each of the four research questions. It is divided into four subsections, which address the results and discussion relating to: 1) CAF measures in direct (OPI) and semi-direct (VOCI) testing modes, 2) CAF measures and task type, 3) Test takers' perceptions of direct and semi-direct testing modes, 4) Relationship between test takers' perceptions and test performance. The table below explains acronyms that will be used in the chapter.

Table 8.

Descriptions of the Acronyms

Acronym	Description
VOCI	Video Oral Communication Instrument
OPI	Oral Proficiency Interview
ASU	Analysis of Speech Unit
ML-ASU	Mean Length of ASU
MS-TTR	Mean Segmental Type-Token Ratio
EF-ASU	Error-Free-ASU
SP/ FP	Silent pause / Filled pauses
CAF	Complexity, Accuracy, Fluency

CAF MEASURES IN OPI AND VOCI

This section starts with descriptive statistics of the CAF measures. Then, inferential statistics will be presented. More specifically, this section will provide the analyses of CAF measures of four types of tasks: comparative, descriptive, role play, and past tense narration. As for the complexity measure, I used the two grammatical measures (number of ASUs and length of ASUs, which refers to the mean of the total number of tokens divided by the total number of ASUs per 100 words), and the Mean Segmental Type-Token Ratio (MSTTR). In order to calculate the MSTTR, the measures of token counts, type counts and type-token ratio had to be calculated. Then, the means of TTRs are calculated. The reason for using two grammatical measures is that using the number of ASUs alone does not necessarily indicate the complexity of the utterance. For example, the speaker might produce many short utterances and hence achieve a high score even though the utterances are simple. For this reason, this measure is best used alongside mean length of utterance or in this case mean length of ASUs (Ellis & Barkhuzien, 2005, pp.152-154). As for the lexical measure, MSTTR was used because using the TTR by itself is influenced by text length. Ellis & Barkhuzien (2005) suggested high TTR can be achieved much easier in shorter texts than the longer ones. I had to calculate the tokens and types in order to measure the type-token ratio for each segment of 100 words, then the mean scores for those segments were calculated. As for the accuracy measure, percentage of error-free ASUs was used, which is one of the common and general measures of accuracy (Ellis & Barkhuzien, 2005). Fluency was measured in

terms of the silent pauses (SP and filled pauses, where the former refers to any pause greater than 400ms, and the latter refer to the usage of the gap fillers “ah” and “like.”

This section begins with the descriptive statistics for the CAF measures, followed by the inferential statistics. After that, findings of the complexity measures will be discussed, followed by the accuracy measure, and finally the fluency measure. Examples for each measure will be illustrated.

Table 9 shows the mean scores of the specific CAF measures, including ML-ASU, MSTTR, percentage of EF-ASU, SP, and FP. The table also shows the four tasks (comparison, description, role play, and past tense narration) that were used for analysis. For complexity measures, participants got the highest mean scores in the VOCI testing modes in the comparative, descriptive, and role play tasks. In regard to accuracy measured by percentage of error-free ASUs, participants had higher mean scores in the OPI in all of the given tasks. With regard to fluency measured by silent and filled pauses, participants got higher mean scores in the VOCI except for the past tense narration.

Table 9.

Descriptive Statistics of Task Type and Complexity, Accuracy, and Fluency (CAF) Measures (N=9).

Task	Lexical Complexity						Accuracy			Fluency					
	ML- ASU			MS-TTR			EF-ASU			Silent Pauses			Filled Pauses		
	Md	M	SD	Md	M	SD	Md	M	SD	Md	M	SD	Md	M	SD
Comparison															
OPI	8.6	9.3	2.8	0.52	0.53	0.1	80	82	10	3	2.9	0.83	4	3.7	1.2
VOCI	15.4	14	3.2	0.5	0.48	0.1	63	63	8	7	6.8	2.3	8	9.3	3.0
Descriptive															
OPI	6.9	6.9	1.6	0.45	0.49	0.1	80	81	15	4	4.6	1.0	5	6.6	3.2
VOCI	13.2	14.9	2.7	0.52	0.68	0.1	70	72	19	10	8.9	3.7	9	8.3	1.9
Role Play															
OPI	8.3	8.7	3.9	0.49	0.48	0.1	70	73	8	8	7	2.8	6	6	2.6
VOCI	9.8	10.5	3.9	0.48	0.5	0.1	60	60	7	10	9.3	3.3	7	7.6	2.2
Narration															
OPI	17	17.1	4.9	0.68	0.48	0.1	89	87	10	8	7.2	3.5	15	15.8	3.2
VOCI	10.7	13.7	4.4	0.72	0.71	0.1	67	82	23	7	7	3.2	8	8.6	2.6
Total															
OPI	41	43	13	2.14	2.17	0.4	319	559	43	23	22	8.13	30	32.1	10.2
VOCI	49	53.1	14	2.22	2.18	0.4	260	277	57	34	32	12.5	32	33.8	9.7

The results were analyzed and compared using the Wilcoxon Signed Rank test, which is a version of the T-test used for small samples. The reason for using the Wilcoxon test is to test if CAF measures' differences between OPI and VOCI are statistically significant.

Participants had a significantly higher percentage of error-free ASUs on the OPI in all of the four tasks (comparison: M =82 , SD = 10, Md=80; description: M=81, SD=15, Md=80; role play: M=73, SD=8,Md=70; narration: M=87, SD=10, Md=89) compared to the VOCI (comparison: M =63 , SD = 8, Md=63; description: M=72, SD=19,Md=70; role play: M=60, SD=7,Md=60; narration: M=82, SD=23,Md=67), meaning that they were more grammatically accurate, making fewer errors; z=-5.071, $p<0.05$.

Table 10.

Wilcoxon Signed Rank Test

	VOCI-ML-ASU OPI-ML-ASU	**VOCI-MSTTR OPI-MSTTR**	**VOCI-EF-ASU OPI-EF-ASU**	**VOCI-SP OPI-SP**	**VOCI-FP OPI-FP**
Z	-1.906[b]	-.358[b]	-5.071[c]	-3.047[b]	-.693[b]
Asymp. Sig. (2-tailed)	.057	.721	**.000**	**.002**	.488

In the following ranks table, we can also see that 33 responses indicate a negative rank in EF ASU, which means that 33 out of 34 had more errors in the VOCI.

Table 11.

Signed Rank Test

		N	Mean Rank	Sum of Ranks
VOCI-MLASU OPI-MLASU	Negative Ranks	9[v]	20.67	186.00
	Positive Ranks	25[w]	16.36	409.00
	Ties	0[x]		
	Total	34		
VOCI-MSTTR OPI-MSTTR	Negative Ranks	15[j]	17.37	260.50
	Positive Ranks	18[k]	16.69	300.50
	Ties	1[l]		
	Total	34		
VOCI-EFASU OPI-EFASU	Negative Ranks	33[m]	18.00	594.00
	Positive Ranks	1[n]	1.00	1.00
	Ties	0[o]		
	Total	34		
VOCI-SP OPI-SP	Negative Ranks	8[p]	13.81	110.50
	Positive Ranks	25[q]	18.02	450.50
	Ties	1[r]		
	Total	34		
VOCI-FP OPI-FP	Negative Ranks	12[s]	18.92	227.00
	Positive Ranks	20[t]	15.05	301.00
	Ties	2[u]		
	Total	34		

Complexity Measure

The following examples show the analysis of the complexity measure in both OPI and VOCI. I need to mention that in the following examples, I am only showing the complexity measures (ASUs length, and MS-TTR). For that reason, I am not showing the accuracy or the fluency measures. You can see that the filled pauses are included; however, because those words (ah, like) are the participants' productions, I did not delete them as they are part of their tokens. The tasks chosen here are the description and past tense narration. I chose these tasks for two reasons. First, in syntactic complexity,

participants had higher mean in the VOCI in three tasks (comparison, description, and role play) and description has the highest mean (M=14.9, SD=2.7, Md=13.2). Second, I chose the past tense narration because MS-TTR was higher in three tasks in the VOCI (description, role play, and narration), with narration having the highest mean (M=0.71, SD=0.1, Md=0.72). The reason for choosing two tasks only is because there were no significant differences in terms of the complexity measures in both testing modes.

In Example 2, Mohamed responds to questions asking him to describe his hometown.

Example 2: Descriptive task (Mohamed)

OPI: Describe your hometown.

My hometown is Dammam, which is located in the {ah} western side of Saudi Arabia/ it is big and {ah} wide city/ it has a lot of {ah} big malls, and many restaurant, and fun places to go/ the weather is {ah} very hot and humid/ it has the big oil company "Aramco"/ many people who live in Dammam work in Aramco/ but not all of course/ it has large diversity/ also we have nice beach nearby the city/ I think that is what I can remember because I am pretty sure there are more to say about my city.

VOCI: Describe your hometown.

My hometown is Dammam/ originally, I am from Alhassa/ but I move to Dammam long time ago/ I honestly consider it my hometown/ it is located in the {ah}

western coast of Saudi Arabia/ it is very popular because it has {ah} biggest and famous oil company in the world/ it is called Aramco/ my hometown is {like} the way .../ it is in the middle of{ ah} the dessert/ its side is known for its palms palm trees world famous for planting dates/ the people as well are {like} the most trained people in the whole kingdom/ it is a great city/ also it is modern and civilized/ folks there are great {ah} people/ they are used to meet people from different {ah} districts/ so race is not a problem there in most cases I would say/ I don't know much to talk about my place/ I would say I wouldn't be able to live away from my city without a comeback.

In example 2, the ML-ASU in the OPI was (10.00) while in the VOCI the ML-ASU was (9.90).

In Example 3, Ahmad performs the past tense narration task, which is analyzed in terms of the MSTTR per 100 words.

Example 3: Past tense narration task (Ahmad)

OPI: Can you tell me an unforgettable experience that you had?

So, an unforgettable experience is when I went with my brother and my sister in law and my nephew when we went to San Antonio/ {ah} it was a road trip/ and it was really fun because {ah} we.... got to... I went to ah six flags/ I enjoy the rides and{ah} enjoy everything/ that was the first on the list for me so that was a lot of fun/ I rode every single one of the games / I enjoyed/ my brother enjoyed it/ and my sister in law enjoyed it as well/ then I went to... wait ... {ah} they had there it wasn't SeaWorld there were there was that in L.A/ I can't remember/ sorry I was

just kind of all over the place now/ but I remember the restaurants were from the City-Walk/ {ah} {ah} {ah} what was also nice/ I like the river over there/ San Antonio is famous for the river that is called river walk/ then, we visited the Mexican market, which is amazing/ they have their own dresses, {ah} kitchen stuff, and {ah} decorations as well/ I bought a lot of {ah} things from them/ after that, we went to the {ah} outlet/ they have very {ah} {ah} huge outlet/ ah {ah} I forgot what is it called/ I think that is what I can remember/ it is not the places that made it a special experience/ but it is the {ah} the {ah} the company with my brother and his wife.

VOCI: …… have you ever had such an experience—an experience that you'll never forget….. It can be something positive or it can be something negative…..Tell us about it.

One of the experiences I cannot forget is {ah} when I move to Stillwater from Saudi Arabia/ I use to live in a big and {ah} {ah} let's say like busy city/ I use to go out at midnight with friends/ and every weekend we go to the beach, like almost very weekend / when the first day I arrived, it was night/ so I did not see much of the town/ in the morning, I went to see the city/ I was shocked/ I honestly felt like depressed the first few days/ I was like {ah} was like what I am going to do here/ I tried to apply for other universities/ but it is {ah} {ah} hard to find admission/ when I met the Saudi community ah, I was like OK that is a good group/ but {ah} now I feel attached to {ah} this place/ and {ah} I have wonderful memories with my friends.

In the OPI, the MSTTR is 0.60 (TTR of 0.55 for first 100 words, 0.65 for second 100 words) while in the VOCI it is 0.61 (total response length was 100 words).

Although there is no significant difference between OPI and VOCI in terms of the ML-ASU and MS-TTR (z=-.358, p=.721), some participants have higher MS-TTR and ML-ASU in the VOCI (See Table 9 for descriptive statistics). I believe that the bouncing balls showing time remaining on the VOCI could have encouraged participants to produce more tokens before the balls disappeared. Whereas on the OPI, participants just said what they had in their minds and once they were done, the interviewer asked them the next question.

Accuracy Measure

Unlike complexity measures, accuracy measures showed a significant difference between testing modes. In examples 4-7 below, participant responses are analyzed for accuracy (EF-ASU).

Example 4: Comparative task (Aziz)

OPI: compare your hometown to the capital.

*My hometown is Jeddah/ and it is in the western region of Saudi Arabia / {ah} it is famous for being located in (**grammatical**) the red sea {ah} {ah} coast/ it is very busy and alive/ there are {ah} {ah} many places to go and have fun with your family/ because it is near the red sea, it is very humid and very hot in the summer/ {ah} in the winter, the weather is good because it is not very cold/ {ah} it is just cool weather at that time/ it never gets very cold/{ah} it is one of tourist cities in*

Saudi Arabia/ people come to visit Jeddah during the summer and other holidays/ ah I think it is very beautiful and always alive/ I mean there are always activities, carnivals and celebrations, yeah

VOCI: can you compare your hometown with a city you visited or know very well?

I will compare Jeddah to Stillwater/ {ah} first of all the two places are way way different/ but there might be some similarity ***(grammatical)****/ {ah} first Jeddah is very big/ and it has different districts and areas/ but Stillwater is like one district in Jeddah/ {ah} it is very small/ Stillwater is just for college/ there is nothing else here/ but in Jeddah there are a lot of thing (****grammatical****) to do besides educational places/ the location of the two cities are (****grammatica****l) different/ Jeddah is in the west side of Saudi Arabia/ but Stillwater I think {ah} if* ***(grammatical)****not mistaken in the southwest of USA/ the weather might be similar/ both places are hot/ but Jeddah is humid/ and Stillwater is dry/ both places has (****grammatical****) people from different parts of the world/ {ah} / I think Jeddah is more like a big and busy city/ but Stillwater is like a town,*

In Example 4, Aziz's EF-ASU is 88% on the OPI (7 error-free ASUs out of 8 total) and 71% on the VOCI (10 error-free ASUs out of 14 total).

Example 5: Descriptive task (Talal)

OPI: Describe a city that you visited

One of the best places I have visited is Madrid in Spain/ I have visited many places/ but {ah} Madrid is one of the very special places/ I loved the atmosphere of the city/ it is just amazing/ I went to the famous museum/ I think {ah} it is called ah the Prado/ it is like {ah} a place that has collections of masterpieces of unique pieces/ those pieces are like from the very {ah} old times/ I also visited several parks, malls, and most importantly the stadium/ it is called Santiago/ as {ah} ***(grammatical)*** *fan of Real Madrid, it was like one of the must go for me/ the city itself is special/ there are so many things I did in ah in there/ I cannot recall all of them*

VOCI: Describe one of your best friends.

I really have a lot of {ah} best friends, not just one/ OK/ I will pick up one of them/ {ah} {ah} my friend Amro/ he is my neighbor as well/ he is like a funny guy/ everything is easy and possible for him/ {ah} our families knows ***(grammatical)*** *each other since we are neighbors/ we {ah} {ah} do not share anything together/ {ah} {ah} I mean we are like like very different personality* ***(grammatical)****/ but we work well together/ {ah} {ah} I like him because he is {ah} {ah} very how to say it {ah} , very dependable/ I think that is the right word/ he is there for me whenever I need him/ he has {ah} a very kind heart/ he never gets mad on* ***(grammatical)*** *anybody/ he is like really a cool guy*

In Example 5, Talal's EF-ASU is 92% on the OPI (11 error-free ASUs out of 12 total) and 83% on the VOCI (10 error-free ASUs out of 12 total). This performance follows the general trend of being more error-prone in the VOCI.

Example 6: Role play task (Salman)

OPI: You have been working in Stillwater to promote multicultural awareness and you have won an award as the person who made the most contribution to multiculturalism in Stillwater. We are at a luncheon in your honor. And. You need to make a very brief acceptance speech for this award. I will introduce you and clap and then you can make your very brief speech.

*Hello, everyone/ {ah} thank you for choosing me to represent everyone of you/ it means a lot to me/ I do accept the award/ and {ah} it is a great honor for me/ I want to remind you of the nice diversity we have in Stillwater/ and I hope that it continue (**grammatical**) to be that way/ thank you again/ it is a pleasure to be one part of this community/ being an international student myself , {ah} I can see the advantages of promoting for the multicultural awareness/ one of the advantages is to create a good atmosphere for people from different parts of the world/ so they feel welcomed and involved/ let's keep the good work up (**grammatical**)*

VOCI: You have a summer job selling great books, I am a potential customer, convince me why I should buy the books from you.

*Hi sir/ do you want to have a look at my books/ {ah} I have some great one (**grammatic**al) / I have (**grammatical**) best offers in town/ {ah} you know what/ I do have sales during the summer/ so if you enjoy reading, this is {ah} (**grammatical**) best time to buy books/ I guarantee that you will not find cheaper books than me (**grammatical**)/ {ah} {ah} I have also good reviews/ go to google/*

and find my store/ and you will not find a single bad review/ I always try to please my consumers/take your time/ and go through the books I have/ and I will give you the best price you could afford

In Example 6, Salman's EF-ASU is 83% on the OPI (10 error-free ASUs, 12 total) and 71% on the VOCI (12 error-free ASUs, 16 total).

Example 7: Paste tense narration task (Alali)

OPI: Can you tell me an unforgettable experience that you had?

{ah} It is a good experience/ Last year, I traveled with my friends to Turkey, Istanbul/ it was my first visit to {ah} Turkey/ we were like ah group of five/ the city is {ah} amazing/ {ah} {ah} it is very beautiful/ {ah} I remember the first day when we visited an island called {ah} " princesses island"/ I like it because it is {like} an untouch ***(grammatical)*** *island/ cars are not allowed there/ people use horses, and bikes only/ it is {like} a tourist city/ so nobody lives there/ then, we {ah} visited different districts in Istanbul city, like malls, {ah} restaurants, {ah} farmers market, {ah} and museums/ we use (****grammatical)*** *to go out at night and enjoy their Turkish tea on the {ah} sea ports/ we also use (****grammatical)*** *to go watch dancing and {ah} {ah} fun activities/it was{ like} a special experience.*

VOCI: …… have you ever had such an experience—an experience that

you'll never forget….. It can be something positive or it can be something negative…..Tell us about it.

*An unforgettable experience is when {ah} I lose (**grammatical**) my friend in a car accident/ it was {ah} {ah} very hard for me to accept his death/ we were together the day before the accident/ the following {ah} day I heard he die (**grammatical**)/ I was {ah} like shocked because I never thought I would lose him/ until today when I {ah} passed (**grammatical**) by his house, I always remember him and I pray for him/ you know I still have his phone number in my phone/ I do not know why/ but I still have it/ I really cannot forget that day/ there are definitely other experiences that I cannot forget/ but that is one of the hardest*

In Example 7, Alali's EF-ASU is 92% on the OPI (12 error-free ASUs, 13 total) and 70% on the VOCI (7 error-free ASUs, 10 total).

The majority of the participants had higher accuracy in the OPI. As non-native speakers of English, it is possible that when the participants were talking to a human being, they tended to monitor their speech and make an effort to avoid making errors. O'Loughlin (1995) believes that certain CAF measures are affected by test takers' perceptions of the time when their performance will be evaluated. It seems plausible that my participants are aware that in the VOCI, their performance will be evaluated or analyzed later; however, in the OPI, participants know that their performance was being evaluated by a native speaker and at the same time while doing the interview. O'Loughlin (1995) talked about the test takers' perceptions of when their performance will be assessed. He further claimed that, in the tape-based tests, testees know that their performance will be assessed later, not at the same time of taking the test. He stated (p. 236)

candidates are clear that their communicative goal is to create a record of their performance for raters displaced in time and space. In live tests, however, it is not always apparent when the assessment will occur. It is possible in the live version of this test that candidates assumed the assessment was being carried out at the time of the test.

Fluency Measure

Fluency measures also showed a significant difference between testing modes. Although there was no significant difference in filled pauses between testing modes, the difference in silent pauses was significant (z=-3.047, p=0.002), (see Table 9 for descriptive statistics). However, no significant differences were found in terms of the filled pauses (z=-0.693, p=0.488), (see Table 9 for descriptive statistics). In examples 8-11 below, responses are analyzed for fluency by counting silent pauses (longer than 400ms) and filled pauses (using "ah" and "like").

Example 8: Comparative task (Salman)

OPI: compare your hometown in Saudi Arabia to any other city.

I am gonna compare my hometown to well let's say {ah} Houston. I have visited Houston a lot/ [>400ms] {ah} my hometown Dammam looks a little bit like Houston/ many people live there/ it is {ah} a busy city, and {ah} crowded/ {ah}[>400ms] but the thing that is different is {ah} the lifestyle there/ {Like} people[>400ms] have different culture different lifestyle, different ways of spending their time/[>400ms]{ah} people in Dammam they usually usually hang

out together a lot/ so let's say {ah} they go {ah} far from the city and hang out there/ {like} they spend their time away from the downtown, away from malls, away from where they actually live/ {ah} {ah} Houston is different/ they usually go downtown to spend their time in the bars, or night clubs, or [>400ms] anywhere else in downtown/ yeah life style is different

VOCI: can you compare your hometown with a city you visited or know very well?

There are a lot similarities between my hometown and the cities I have visited/ so [>400ms]{ah} Let's say for example the similarity similarities and different similarities between my hometown and Stillwater/ Well the similarities are that people in Stillwater are young {ah} because of the college, because it is {like} a college town/ and {ah} most of its residents are students/ and there are a lot of young people in my hometown [>400ms] {ah} because we are a relatively young country/ and there are some differences as well between my hometown and Stillwater/{ah} so the differences are the culture here is different/ for example, the festivals the holidays here and back home are different/ {ah} for example here they have {ah} a lot of people here celebrate Thanksgiving/ [>400ms] back home people don't celebrate Thanksgiving/ people celebrate Eid Alfitr which for Muslims which is an Islamic holiday so those are pretty much the similarities and differences between the both cities

In Example 8, Salman had 4 silent and 11 filled pauses in the OPI, compared to 2 silent and 6 filled pauses on the VOCI.

Example 9: Descriptive task (Odai)

OPI: Describe your hometown

Well my hometown is {ah} farm-based city/ it is like a town/ [>400ms]It is based on agriculture/ So my home city is based on farming right/ So {ah} [>400ms]I consider it to be a green area/ a lot of people used to work on the farms until they opened Aramco, which is the oil company/{ah} so[>400ms] a lot of people just got into that business{like} working for Aramco/ [>400ms] so we switch from farming to that oil business/ So {ah} farming now is being less popular than it used before/ but {ah} [>400ms] it is still known for that/ I mean that is the thing i can tell you about it/it has high population/ {ah} it is crowded/[>400ms] it does not {ah} have good public transportation because people there use their cars/ {ah} oh by the way it is called Al-hassa/ it is on the Eastern part of Saudi Arabia/ I forget to mention that/ that is wired/ {ah} yeah I would just keep it to that/ the list goes on and on/ but I just keep it to that

VOCI: Describe one of your best friends.

Yeah to describe this one guy/ that is my best friend/ We have been competing in about [>400ms] about 10 years/ that is when we started competing with each other/ So {ah} whenever I get something, he would get the other/ We would compete on grades {ah} {like} university admissions/ / So [>400ms] one thing I liked about him that he {ah} does he knows how to do stuff perfectly/ You know he does he does his thing/ I like him/ and so I always refer to him as a doer/ he knows how to do stuff / and I like the way we have been competing because it

pushed me to do better/ and {ah} that push him to do better as well/ and I like that attitude/ {ah} and that he still keeps it professional although sometimes it is intense/ but yeah

In Example 9, Odai had 5 silent and 7 filled pauses in the OPI, compared to 2 silent and 4 filled pauses on the VOCI.

Example 10: Role play task (Alabad)

OPI: You have been working in Stillwater to promote multicultural awareness and you have won an award as the person who made the most contribution to multiculturalism in Stillwater. We are at a luncheon in your honor. And. You need to make a very brief acceptance speech for this award. I will introduce you and clap and then you can make your very brief speech.

thank you everyone for the award/I am so happy today that I am among you/[>400ms] and I want {ah} to encourage and {ah} {ah} support you to accept diversity and respect people [>400ms] of different colors, religions, and races and regardless of what they believe in/ /[>400ms] I worked hard to {ah} promote multicultural awareness because {ah} {ah} Stillwater has people from different places/ so[>400ms] it is important to make people feel respected and welcomed /[>400ms]I think multicultural awareness is important because that contributes to make the place more productive and more united/{ah} [>400ms]/ when we are united, we become stronger/ and we help each other to build a healthier community/ so it my pleasure to be part of you/ and {ah} I hope we continue being tolerant and open to other cultures and other differences

VOCI: You have a summer job selling great books, I am a potential customer, convince me why I should buy the books from you.

Hi guys/ do you want to come and have a look at my books over here/ {ah} I have great selections and excellent sales/[>400ms] I also have good offers/ {ah} you can buy one, and get the other {like} {ah} half price/ if you buy two, get one free/ I am sure you like the {ah} great selections I have/ [>400ms] I have never had any costumer {like} {ah} come back to me because they all love my books/ I also {ah} {ah} I have some old books that are {ah} rare/ you cannot find them easily/ and I have new books that are just like {ah} new in the stores/ {ah} If you are a student, I can do student discount/ you can take few minutes and browse through the books and see if you like it/ I have a small reading booth/ you can take a book and go to the booth and read few pages/[>400ms] if you tell me what you like, I can at least help you to choose.

In the previous examples, Alabad had six silent pauses in the OPI and two silent pauses in the VOCI.

Example 11: Paste tense narration task (Talal)

OPI: Can you tell me an unforgettable experience that you had?

{Ah} [>400ms] well I like theme parks right/ So one of the experiences that

I would never ever forget is that one of these rides [>400ms] that is I guess I do not know what is it called/ but they do they attach you to a rope and they pull you up. I would say it is more than 30 meters/ I would say which is about 90

feet/[>400ms] that was really high distance/ [>400ms]and they just {like} once you go up, they ask you to release yourself and you do the release/ so when you release it, you fall really fast just free falling/ that is what you do/ {ah} [>400ms] It is really a nice experience that I never forget/[>400ms] when I release myself and fall down, I appreciate the fact that I am still alive/but it was fun/ it was a new experience {ah} I never did before

VOCI: …… have you ever had such an experience—an experience that you'll never forget….. It can be something positive or it can be something negative…..Tell us about it.

yeah I have been/ I went to Tonkawa/ it was a small village/ It is one hour from Stillwater/ So at that time, there was a terrorist attack in France/ So, Tonkawa is a small village/ so everyone there looks at me as I am a terrorist/ and one of the guys asked me are you one of them/ I was shocked that he asked me that question/ I was very intimidated/ {ah} [>400ms] just because I have a darker skin color and I look different from them, they {ah} think that I am a danger/ at that time, I was planning to study a transfer credit from the college there/ but after that incident I changed my mind/ {ah} I did not feel safe being there/ people there are not used to see international students/[>400ms] I was very scared for my life because people in Tonkawa carry guns with them everywhere/ {ah} I have seen so many people with guns/ so I decided to take the course in OSU/{ ah} that was an unforgettable experience for me/{ah} it was not a good one

In Example 11, Talal had 4 silent and 3 filled pauses in the OPI, compared to 1 silent and 2 filled pauses on the VOCI.

As shown in Table 10, the accuracy measure EF-ASU ($z=-5.071$, $p=0.000$) and the fluency measure SP($z=-3.047$, $p=0.002$) differ significantly in the direct and semi-direct testing modes. Participants had a significantly higher percentage of error-free ASUs on the OPI, and significantly more silent pauses on the VOCI. The findings regarding the fluency measure do not agree with Skehan (2001) and Bygate (2001), who found that dialogic tasks (such as those on the OPI) were produced less fluently than monologic ones (such as those on the VOCI). Nevertheless, it seems that most studies have found dialogue to have faster speech rates, and less pausing, than monologue (Michel, 2011; Riggenbach, 1989; Kowal, Wiese, & O'Connell, 1983; Ejzenberg, 1997, 2000). I also tend to agree with Brown, Cox, and Thompson (2017) who claimed

> The interactive, interpersonal, and synchronous nature of the OPI may have exerted more time pressure on candidates, causing them to quicken their speech rate to maintain the floor, to nominate or change a topic, or simply to avoid silence— a particularly threatening conversational characteristic in the context of an oral exam (p. 804).

It seems logical that participants in the current study wanted to quicken their speech in order to avoid pausing, which could make them lose control over the conversational floor. It is also possible that task type could have affected the CAF measures. This issue will be discussed in the following section.

CAF MEASURES AND TASK TYPE

This section presents the findings and discussion of the relation between CAF measures and task type. Four tasks in each test were used for analysis: comparison, description, role play, and past tense narration. In order to examine the relation between CAF measures and task type, the non-parametric Kruskal Wallis test was run using SPSS 24. In the previous sections, we learned that EF-ASU and SP differ significantly in both testing modes. Now, in this section, we are looking into the association within the testing modes themselves, not between them. For this reason, I used the Kruskal Wallis test in order to find out whether there is an association between task type and CAF measures in each testing mode. Kruskal Wallis was chosen because it is a non-parametric test that determines if the task type has any effect on any CAF measures within each testing mode.

Table 12.

Kruskal Wallis Tests for Association within OPI (N = 34) and VOCI (N = 34) for Complexity Measure

Task	**Syntactic complexity**						**Lexical complexity**		
	Number ASU			**Mean-length ASU**			**Mean Segmental TTR**		
	χ	**p**	**r**	**χ**	**p**	**r**	**χ**	**p**	**r**
OPI	10.809	0.13	0.56	14.698	0.20	0.66	8.476	0.37	0.49
VOCI	1.302	0.729	0.195	4.286	0.232	0.35	10.044	**0.018**	0.54

Table 13.

Kruskal Wallis Tests for Association within OPI (N = 34) and VOCI (N = 34) for Accuracy Measure

Task	**Accuracy**		
	Error-free ASUs		
	χ	**p**	**r**
OPI	7.088	0.069	0.456
VOCI	2.508	0.474	0.27

Table 14.

Kruskal Wallis Tests for Association within OPI (N = 34) and VOCI (N = 34) for Fluency Measure

Task	**Fluency**					
	Silent pauses			**Filled pauses**		
	χ	**p**	**r**	**χ**	**p**	**r**
OPI	9.753	**0.021**	0.54	18.254	.068	0.73
VOCI	4.286	0.232	0.35	1.334	.721	0.19

The previous tables indicate that there is a significant effect of task type on the MS-TTR in the VOCI testing mode (p=0.018, r=0.54), and on silent pauses in the OPI testing mode (p=0.021, r=0.54). Comparing the mean scores for MS-TTR in the VOCI and SP in the OPI, we can see that it is the narration task type that influenced both measures (MS-TTR and SP) for the majority of the participants. The following diagram illustrates the MS-TTR in the four tasks, where the X-axis represent the task type and the y-axis represents the MS-TTR.

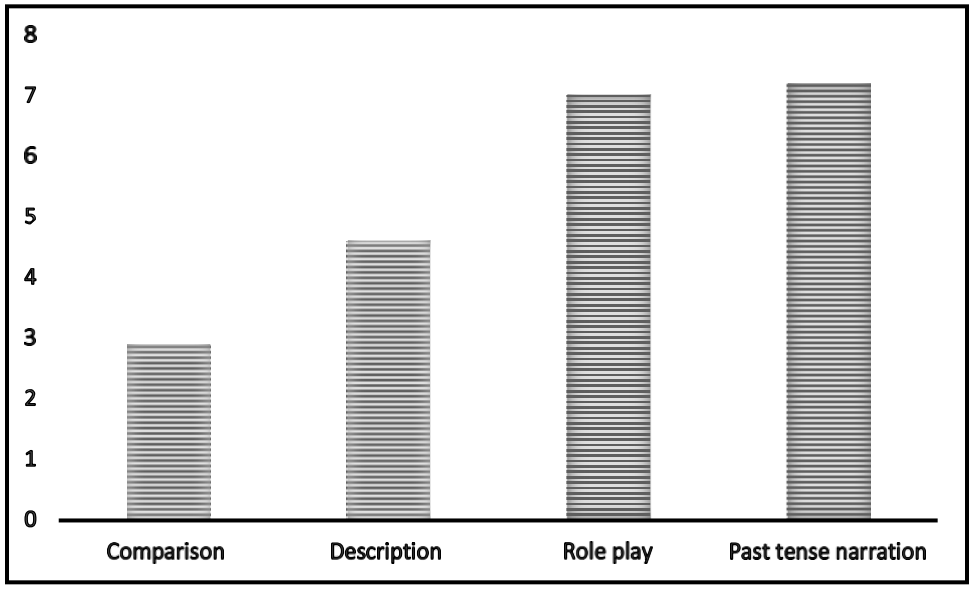

Figure 8. MS-TTR in the VOCI Testing Mode

In Figure 8, we can see that past tense narration has the highest mean, and comparison has the lowest mean. It seems that past tense narration has the highest influence on the MS-TTR. We can look at the following examples, one for the highest mean (narration) and one for the lowest mean (comparison).

Example 12: Past tense narration task (Talal)

VOCI: …… have you ever had such an experience—an experience that you'll never forget….. It can be something positive or it can be something negative…..Tell us about it.

yeah I have been/ I went to Tonkawa/ it was a small village/ It is one hour from Stillwater/ So at that time, there was a terrorist attack in France/ So, Tonkawa is a small village/ so everyone there looks at me as I am a terrorist/ and one of the guys asked me are you one of them/ I was shocked that he asked me

that question/ I was very intimidated/ {ah} just because I have a darker skin color and I look different from them, they {ah} think that I am a danger/ at that time, I was planning to study a transfer credit from the college there/ but after that incident I changed my mind/ {ah} I did not feel safe being there/ people there are not used to see international students/ I was very scared for my life because people in Tonkawa carry guns with them everywhere/ {ah} I have seen so many people with guns/ so I decided to take the course in OSU/{ ah} that was an unforgettable experience for me/{ah} it was not a good one

Similar to what the majority of the participants did, in the past tense narration task in the VOCI testing mode (M=0.71, SD=0.1), Talal had 100 token counts and 55 type counts. Then, the MS-TTR is 0.55.

Example 13: Comparison task (Talal)

VOCI: *can you compare your hometown with a city you visited or know very well?*

I can compare my hometown Jeddah to Vancouver in Canada/ {ah} Vancouver is very cold place, Canada in general is/ ah Jeddah is hot and humid/ {ah) in Vancouver you can see the four seasons/ but in Jeddah it is hot around ah the year/ I think we get some cold weather in winter/ but {ah} we do not get like very cold weather/ {ah} as for population, I think both places has different population/ but the population is different/ {ah} I think the Asian is the largest in Vancouver/ {ah}in Jeddah it is {ah} different because it is more diverse than Vancouver/ we have more ethnic groups in Jeddah / I am not sure/ but that is what I think/ we even have districts for different ethnic groups/ {ah}life style is different/ I think

people are more active in Vancouver than in Jeddah/ maybe nowadays people start being more active

In the comparative task, Talal had 49 type counts per 100 tokens, which makes the MS TTR (0.49).

I have mentioned earlier that the effect of the bouncing ball in the VOCI might have contributed to the higher mean of lexical variation. I believe the bouncing balls have pushed the participants to produce more tokens. One could possibly wonder why the bouncing the balls impacted the MS-TTR but not the fluency measure. One possible explanation is that a person can increase the lexical variation intentionally; however, it is more likely that pausing is done subconsciously.

Now, let's look at the second measure (SP) that was influenced by the task type. In Figure 9, we can see that past tense narration had the highest mean and comparison had the lowest mean score for silent pauses in the OPI testing mode.

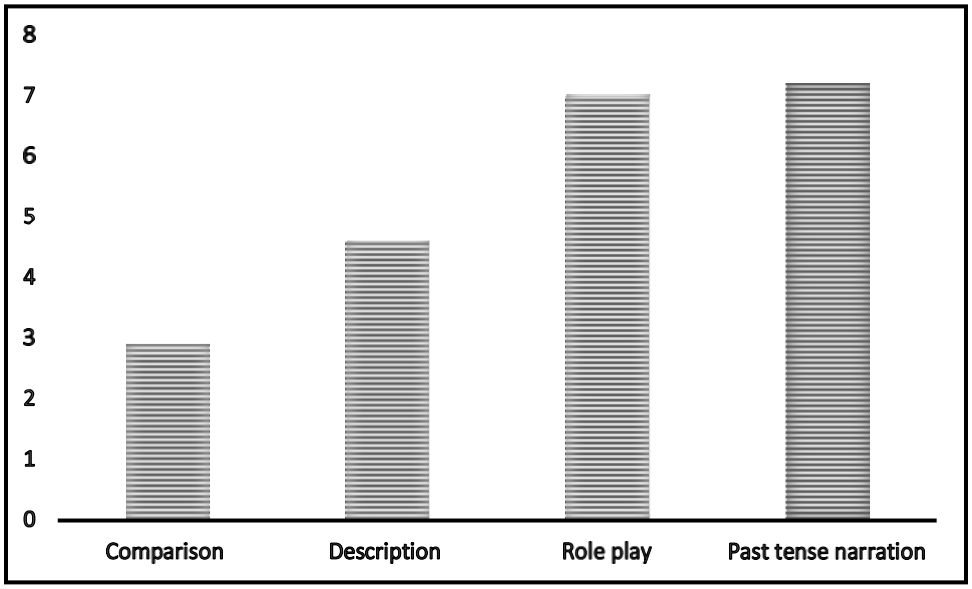

Figure 9. SP in the OPI Testing Mode

Example 14 illustrates Nasser's OPI response to the past tense narration task. Nasser had a total of seven silent pauses per 100 words.

Example 14: OPI past tense narration (Nasser)

Interviewer: Can you tell me an unforgettable experience that you had?

{Ah} [>400ms] I do not really have {like} something specific/ [>400ms] {ah} maybe when {ah} I bought my first car/ it was like {ah} [>400ms] unforgettable because {ah} I just graduate from high school and did not have a car before[>400ms] / it was {like}[>400ms] {ah} a surprise because my parents got it for me/ {ah} [>400ms] my father told me if I get high grades in high school, he will get me a car/ this is {ah} {like} a trend or habit in Saudi/ [>400ms] I mean when {ah} students graduate from high school, their parents get them cars/ I think because because {ah} [>400ms] maybe they will go to college and a car/ {like} you are more mature / {ah} not all parents but the majority/ [>400ms] I had a car just for me/ before {ah} I was sharing cars with my father/ I started driving in high school/ [>400ms] I had to ask permission to drive and if my parents agree/ I take my father's car/ he {ah} will call me like 10 or 11 times to ask me {ah} about where I am/ so it was {ah} a special experience because having a car at that age is something [>400ms], which is different

Example 15 illustrates the same participant's response on the OPI comparison task. He had fewer silent pauses than he did on any other OPI task.

Example 15: OPI comparison task (Nasser)

Interviewer: Compare your hometown to the capital.

The population in the capital is higher/ I guess my city is divided by two sections; the modern section and the old section/ the modern section is {ah} the industrial section/ but the capital does not have that division/ it has some kind of [>400ms] really high tech city [>400ms]/so everything is fully functional by technology, lights/ for example, internet is everywhere from the beginning of life from the start of the city to the end, which is different from the capital because internet is not everywhere there {ah}/ I think also my city has cleaner and wider streets/ one difference maybe my city has more diversity than the capital, I think

In the past tense narration, Nasser had seven silent pauses while in the comparison task he only had 2 silent pauses. This shows that Nasser was less fluent in the narrative task, yet more fluent in the comparative task. The narrative task type did impact the number of the silent pauses in the OPI testing mode. It is possible that the narrative task is different from the other tasks in certain aspects.

Labov (1997) claims that narrative requires speakers to occupy "more social space than in other conversational exchanges - to hold the floor longer, and the narrative must carry enough interest for the audience to justify this action" (Reportability, para. 1). He also emphasizes (1997) that narration about personal experience is different from other narrative tasks, as speakers are often relating experiences that are "emotionally and socially evaluated, and so transformed from raw experience" (Narratives of personal experience, para.2). I think the participants in the current study might have wanted to

occupy more space in talking about their unforgettable personal experiences that they actually have been through and that makes it plausible to claim that they produced more tokens because they were emotionally and socially engaged in talking comfortably about their experiences. I intended to use the term "comfortably" because, as we learned before that the effect of the narrative task on the MS-TTR is significant only in the VOCI, participants felt more secure to talk about their personal experiences in front of the computer, without thinking about their grammar or their language in general.

As for the silent pauses, I think it might be related to the cognitive processing load, proposed by Skehan and Foster (1999). I think the question about "an unforgettable experience" requires a higher processing load because the participants were not only remembering their experiences, but also telling those stories, describing the setting and the emotions involved. I assume that higher processing load in the narrative task could possibly make the participants pause to remember more details. Going back to Labov's (1997) argument about social space, I think the participants had a higher number of pauses in the OPI because they wanted to hold the floor; they did not have to hold the floor in the VOCI due to the absence of the interlocutor.

PARTICIPANTS' PERCEPTIONS/ PREFERENCES TOWARDS DIRECT AND SEMI-DIRECT TESTS

This section presents the findings of the test takers' perceptions towards direct (OPI) and semi-direct tests (VOCI). After analyzing and coding participants' responses to questions about perceptions towards these testing modes in the VOCI and Arabic interview, three major themes (interaction, test structure, and test taker's personal affective factors) and 19 sub-themes were identified. Table 15 illustrates all the themes

and subthemes, as well as the number of participants who mentioned each subtheme. I need to mention here that Table 15 illustrates all the themes found in the data, including the themes that show perceptions towards OPI and towards the VOCI. I also need to emphasize that while some of the subthemes were mentioned only once by one participant, I believe that they are worth mentioning because they still add up towards to the major theme.

Table 15.

Frequency of the themes of participants' perceptions towards OPI and VOCI

Theme	Sub-themes	Number of participants who mentioned the theme/ sub-theme
1.Interaction	Presence of a human being	9
	Ambiguity when there is no interaction/ explanation/ clarification	9
	Involvement/ Engagement	9
	Comfortable setting	8
	Tailored questions	7
	Examiner's reaction	5
2.Test structure	Authenticity	7
	Formality	5
	Contextualization	4
	Topic Familiarity	3
	Test goal	1
	Test score	1
	Critical thinking skills	1
	Artificiality	1
3.Test takers' personal affective factors	Motivation	9
	Personal style	3
	Consciousness	1
	Boredom	1
	Face image	1

Perceptions towards OPI

In this section, I will present the different themes that indicate the test takers' perceptions towards OPI. The V following the pseudonym stands for VOCI, the A for the Arabic interview; the first number indicates the major theme as numbered in Table 15 and the second number indicates the total number of themes mentioned by the participant. In other words, Talal (V/1/11) means that this quote comes from Talal's VOCI, indicating the first theme "interaction," and the total number of themes found in all of Talal's responses (11).

Interaction

This section shows only the interaction sub-themes related to the test takers' perceptions towards OPI. I need to emphasize here that the theme of "interaction" was mentioned by all participants using different sub-themes (see Table 15). The subthemes do not always have the word "interaction" in them; however, they involve some interactive aspects.

Talal (V/ 1/ 11) reported that he preferred communicating with a human being over communicating with a computer in Example 16:

Example 16:

> *I prefer to have a test with a person face-to-face in order to communicate together, not with a computer.*

Another aspect of interaction is the ability to see the reaction of the interviewer. Mohammed (V/ 1/5) described the effect of the interviewer's reaction to him in Example 17:

Example 17:

> *I was able to get an impression for each question I answered, I mean a good impression that the interviewer is interested in what I was saying.*

Odai (V & A/1/8) emphasized how the OPI questions were tailored based on his response in Example 18:

Example 18:

> في المقابله الاسئله كانت بناء على اجاباتي مثلا لما اتكلمت عن صوفيا الجهاز الألي بعدها سألتني ان ليش بيعطونها الجنسيه السعوديه يعني سألتني عن شي انا مهتم فيه وتكلمت عنه من قبل
>
> *The interview asked me questions based on my answers, for example, I said something about the Saudi robot "Sofia", then she said tell me why they want to give a robot a citizenship? So, she asked about something I am interested in and I have talked about it earlier in my answers*.

Another subtheme of interaction is clarification and explanation of the OPI, which is explained by Aziz (A/ 1/7) in this response in Example 19:

Example 19:

في المقابله، كان في مرونه اكثر من الفيديو لان في المقابله مثلا لما سألتني اني اتخيل اني فزت بجائزه ولازم القي كلمة شكر فانا في البدايه ما فهمت فسألتها تعيد السؤال وعادت السؤال طبعا في الفيديو هذا الشي مستحيل لانك تتكلم مع جهاز يا انك تجاوب السؤال و أنك تقول ما فهمت السؤال

In the interview, I think the interview is more flexible than the video because in the interview when she asked me to pretend that I won a prize, at the beginning I could not get it, so I sked her to repeat the question, and she did, but in the video, this is impossible because it is a machine you either answer the question, or simply say I do not understand.

Kanga (2012) found that his participants "saw the examiner as a sort of catalyst, a facilitator, or a supportive listener that they could trust" (p.50). Similar thoughts were also found in this study. For example, Odai (A/1/8) appreciated the interviewer's willingness to explain things to him in Example 20:

Example 20:

في المقابلة ، كانت الأستاذة التي قابلتني تدعمني ، وشرحتلي الأسئله اللي مافهمتها مثلا السؤال حق لعب الادوار مافهمته فشرحت ، شرحها كان مفيد لأني فهمت الأسئلة وبالتالي قدرت اجاوب

In the interview, the professor who interviewed me was supporting me, she explained the questions I did not understand, for example the question about winning the multicultural award, she explained it to me, her explanation was helpful because I understood the questions so I could answer.

The presence of a human examiner is part of what makes the OPI more interactive than the VOCI. Lazaraton (2002) indicated that "the examiner factor is the most important characteristic that distinguishes face-to-face speaking tests from their tape-mediated counterparts" (p.152). In this case, OPI and VOCI have the same guidelines and question functions; however, the OPI has an interviewer, while VOCI is carried out through the computer. Shohamy (1994) also claimed that "the physical presence of a human interlocutor on the OPI is very likely the cause of language production that is more conversational and intimate" (p.118).

Another sub-theme of interaction is engagement/ involvement. Mohammed (V/1/5) regarded the OPI as more engaging than the VOCI as shown in Example 21:

Example 21:

The interview I did was better because it was more interactive and more engaging

Five participants reported that they were more involved in the OPI than in the VOCI. Alali (V &A/1/10) described how involved he was during the OPI as in Example 22:

Example 22:

عكس في المقابله لاني كنت جدا مستمتع بالمحادثه والحوار

"…….*Unlike the interview, I was very much involved in the conversation*.

Alali was asked about what he meant by the fact that he was "involved in the conversation" and he elaborated in Example 23:

Example 23:

انا كنت اقصد اني مندمج في المحادثه حقت المقابله يعني انا كنت جزء من المحادثه لان الدكتوره سألتني اسئله وكانت تسمع لاجاباتي وتعلق عليها هذا اللي خلاني اكون انفولفد وجزء من المحادثه بعكس الفيديو ماجاني هذا الاحساس لان كانت محادثه من طرف واحد

What I meant by being involved in the conversation during the interview is that I was part of the conversation because the professor asked me questions, listened to my answers, and also commented on some of my responses, so I felt I am involved and I am part of this conversation, unlike the video because I did not feel that way because it was just one-way interaction (personal communication, October 27,2019).

Through the participants' responses, we can conclude that there are some possible reasons for them to feel more involved in this testing mode. For example, participants mentioned that they like the communicative nature of OPI and the presence of a person. I think that the interviewer's presence, and possibly her interviewing strategies made the participants feel involved. For example, interviewing strategies were mentioned by Nasser (A/ 1/13) in Example 24:

Example 24:

بعض الا شياء اللي في المقابله ان الدكتوره كانت مثلا تهز راسها دليل عل انها تتفق معي ممكن او كانت تبتسم وهذا دليل انها مستمتعه بالاجابات

Some of the things the doctor did in the interview that made me feel more involved is that for example she nodded her head which means to me she agrees with what I say, maybe, also she smiled and that tells me she likes my answers.

Talal (A/1/11) agreed in Example 25:

Example 25:

كان فيه بعض السيقنالز اللي خلتني احس بالاندماج في المقابله مثلا كانت تقول واو قولي اكثر عن هذا الموضوع مع]ان اعرف ان جوابي عادي بس ممكن من باب التشجيع واتذكر كمان كانت تقولي اوه انتريستنق فهذا يعطيني شعور ايجابي مشجع

There were some signals she used that made me feel more involved in the interview, for example, she said "wow," "tell me more about it," and I know that my answer is not very special but that is very encouraging, I also remember when she said "oh interesting," and that gave me a positive and encouraging feeling.

Ahmad (A/1/8) provided details about how the interviewer showed interest in his responses in Example 26:

Example 26:

اللي سوت معي المقابلة كانت لطيفه معي وكانت مستمعه جيده ماكانت بس تسمع عشان تشوف جوابك صح او غلط لا كانت تبين لي انها مستمتعه باجاباتي و كانت تخليني اقول تفاصيل اكثر اتكلمت عن تجرتي في سكاي دايفينق وسالتني اسئله كثيره عن تجربتي وهذا اسعدني انها مستمتعه بالشي اللي قاعد اقوله وكانت مهتمه تعرف تفاصيل اكثر

The interviewer was very nice and she was a very good listener who does not only listen whether you answer or you did not, but she showed interest in what I was saying and she really wanted me to say more details, I talked about my experience

of sky-diving and she asked me many questions about my experience, which made me happy to know that she is interested in what I was saying and she cared about knowing more details

Many participants viewed OPI as more comfortable than VOCI. Participants' responses about this theme indicate that it is the examiner who made this testing mode more comfortable than the VOCI, and that explains why I added this sub-theme under interaction. Ahmad (A/ 4/ 8) stated in his response in Example 27:

Example 27:

عشان هذا السبب انا اشوف المقابله مريحه اكثر بالنسبه لي

……*for this reason, the interview was <u>more comfortable</u> for me*.

One could imagine that the presence of a person could create more pressure for test takers; however, eight of the participants reported that the interviews were more comfortable for them. It is very interesting to note that while Alabad compared his testing experience of the VOCI to that of the TOEFL, he views the TOEFL as a comfortable testing mode because there is no interviewer who can be biased against him. Alabad was the only participant who said that in terms of tests he had taken, he preferred both a semi-direct test (TOEFL) and the direct interview (OPI). He was contacted by email and asked to elaborate about this paradoxical opinion. Alabad replied in Example 28:

Example 28:

ايوا مشكلتي في اختبار الايلتس هو الاكزامنر لان تسمع قصص عن ايكزامنرز عنصريين بس عشان شكلك او عرقك او دينك بالنسبه لي ذا اقدر اضمن ان الايكزامنر عادل ماعندي اي مشكله بالنسبه للمقابله البروفيسوره اللي

سوت معي المقابله كانت جدا فريدنلي وكويسه ف انا ماعندي اي مشكلهاذا يختبرني ايكزامنر زيها لكن للأسف لما تختبر ايلتس ماتقدر تعرف من اللي بيقابلك

Yes, my main problem with the IELTS exam is the examiner himself because you hear stories about examiners being biased against you just because of how you look like or your race, so for me if I can guarantee that the examiner is fair enough to me, I have no problem with that, however, in the interview I did, the professor who interviewed me was very nice and friendly, so I do not mind having the IELTS with an interviewer like her, but unfortunately, when you have IELTS, you never know who is going to interview you. (personal communication, October 17, 2019).

Aziz, too, viewed OPI as more comfortable than the VOCI as shown in Example 29:

Example 29:

في اختبار الفيديو كنت متوتر لان كنت احاول اركز عشان اسمع السؤال زين لان لو ماسمعت السؤال خلاص ضاع عليك لكن في المقابله كنت مرتاح و ريلاكسد لان حتي لو ماسمعت السؤال زين او مافهمته اقدر اسأل اللي يختبرني يعيد السؤال

In the video, I was stressed because I was trying to concentrate in order to listen carefully to the question because when you do not hear the question, then you miss the chance to answer. However, in the interview, I was more comfortable and relaxed because if I did not hear the question very well, or even understand

it, I can always ask the examiner to repeat the question (personal communication, October 19, 2019).

According to Fulcher (1996), “If recording equipment is to be used during the test, its position and proximity to the students must be considered carefully” (p.32). In the OPI there was a tape recorder sitting on the table, but the students were focused on communicating with the interviewer (face-to-face). With the VOCI, there was no person to respond to. That focused the students' attention on the tape recorder, which, possibly, could have caused some anxiety on the part of the students.

Test structure.

This theme refers to any sub-themes that are related to the tests themselves (OPI, VOCI). Authenticity and time were two of the test structure related sub-themes that are used to show perceptions towards OPI. Participants viewed OPI as authentic because it resembles conversations they experience in daily-life, as Ahmad (A/ 2/8) exaplained in Example 30:

Example 30:

في المقابله انت تتكلم مع شخص قدامك زي اللي نسويه في حياتنا اليوميه يعني احنا نتكلم مع اشخاص حقيقيين

In the interview, you are talking to a person in front of you and this is similar to what we do on daily basis, which means that we are talking to real people.

Many researchers have claimed that the OPI is different from natural conversation (Johnson, 2000; Johnson & Tyler, 1998; Van Lier, 1989; Young & Milanovic, 1992). Kitajima (2009) stated that, based on Van Lier’s description of OPI and natural

conversation, “the OPI fundamentally differs from natural conversation, exhibiting an asymmetrical contingency, where one party controls the interaction by initiating, sustaining and terminating talk through a typical question–answer format” (p.146). However, Halleck (2005) asserts that while the OPI interaction might not be conversational, this does not invalidate it as a test of oral proficiency. Despite some researchers’ contention that the OPI does not represent conversation, some participants in this study viewed OPI in a positive way because they felt that it was similar to natural conversation, even though in reality it more closely represented the typical characteristics (and asymmetrical roles) of an interview. According to Moder and Halleck (1998), the OPI is not actually an informal conversation, “but it does sample the communicative behavior of interviewees in an authentic speech event” (p. 144).

Previous research suggests that test takers who view oral proficiency interviews as conversation are more successful than those who do not. For example, Jenkins and Parra (2003) asserted that “participants who framed the interview as a discussion or conversation among peers were more successful than those who framed it as an examination” (p.90).

Time is another sub-theme of test structure. Alabad (V/ 2/13) reacted to the length of the VOCI in Example 31:

Example 31:

> *I felt that the <u>time</u> in the interview was shorter, I am not sure if it is in reality shorter, but the longer time made me more nervous or stressed.*

It is true that the VOCI was 15 minutes longer than the OPI. One could possibly argue that the shorter the test, the more stressed the test taker, as the participant always thinks about the shorter time they have for the task. However, Alabad's above quote contradicts this possibility. This finding does not agree with Suryaningsih's (2014) finding that "participants argued that the time given in the test was too short. It made them feel pressured and stressful" (p.31). One could possibly argue that if short test time creates stress and pressure on test takers, would not a long test time make them more comfortable and assured that they have sufficient time to develop ideas and be more creative with their responses?

Many participants reported that they were more motivated during the OPI than the VOCI. Salman (V/2/8) in Example 32 claimed:

Example 32:

I feel more <u>motivated</u> to talk in the interview I did last week.

When Salman was asked why he was more motivated to talk during the OPI than during the VOCI, he answered (in Example 33):

Example 33:

ايش اللي يحمس الواحد يتكلم لجهاز قدامه اقصد يتكلم ان انت تقول شي والشخص اللي قدامك يرد في المقابله كنت متحمس لان ابغا اقول رأيي في بعض الاسئله واشوف ايش رأي الايكزامنر في اجاباتي او حتي اشوف ردة الفعل

Why would someone be motivated to talk to a computer? And to talk here means to have a conversation, you say something, and the person in front of you replies, I was motivated to talk because I wanted to say my opinion about some questions and see what the interviewer would say about it, or even see the reaction (personal communication, October 19, 2019).

Test takers' personal affective factors.

This theme includes any sub-themes that are related to test takers' personal factors, including motivation, personal style, and consciousness. It can be implied, from participants' responses, that some factors could make participants more motivated during the OPI. For example, participants might want to show that they have good command of the English language. In the examples below, three participants explain how taking the OPI had been a special experience, as none of them had ever spent so much time outside of class getting individual attention from a professor:

Aziz (A/3/7) commented how special his experience was in Example 34:

Example 34:

اول مرا اجلس اسولف مع دكتور غير وقت المحاضره خصوصا هنا في امريكا لان بعض الدكاتره يحسب الطلاب الانترناشونال ما يفهمون انجليزي كويس

It is the first time to sit with a doctor outside the lectures, especially here in America because some doctors think that international students do not understand English very well.

Salman (A/3/8) agreed in Example 35 that for him the OPI represented a special experience:

Example 35:

ماقد حصل وجلست مع بروفيسور نتكلم عن اشياء غير الاختبارات والدرجات او المنهج نفسه

It did not happen before that I sat with a professor to talk about other things other than exams, scores, or the subject itself.

Mohammed (A/3/5) also agreed (in Example 36) that he had never spent time outside of classes with a professor:

Example 36:

بالنسبه لي اول مرا اجلس مع دكتور برا الكلاس ونتكلم عن شي غير الدراسه تجربه حلوه كانت

For me, it is the first time to sit with a doctor outside the class and talk about something other than school subject, it was a good experience

Consciousness is also one of the sub-themes of personal affective factors mentioned by Nasser (A/3/12) in Example 37:

Example 37:

من ضمن الاشياء بالنسبه للمقابله ان كانت زي الفيديو في تسجيل للاجابات ولكن في المقابله انا نسيت موضوع التسجيل

One thing about the interview is that it was recorded just like the video test, but in the interview, I was not <u>conscious</u> about the recorder…"

In a follow-up interview, Nasser was asked whether consciousness about the recorder is positive or negative. His reply appears in Example 38:

Example 38:

اتوقع ان احساسي كان سلبي لان لما تعرف ان كلامك يتسجل تحاول انك تتكلم بطريقة صحيحه وماتسوي اغلاط قرامر او كلمات تستخدمها بطريقه غلط

I think how I felt was negative because when you know you are being recorded you want to make sure you speak good English and not make grammatical mistakes and inappropriate words' usage." (personal communication, October 19th, 2019).

One of the affective factors that influenced test takers' preferences towards the direct and semi-direct testing modes is personal style. Aziz (A/3/7) thinks that OPI suits his personal style, as he explained in Example 39:

Example 39:

عجبتني المقابله لان تناسب ستايلي انا عموما ما ارتاح كثير للتعامل مع الكمبيوتر اعرف كيف استخدمه اكيد بس ماحب استخدمه خصوصا لما يكو ن اختبار عالكمبيوتر

I like the interview I think because it is fit my style, I am not very comfortable with using the computers, I know how to use it of course, but I do not like using it especially for exams.

Personal styles do not only apply to language learning, but also to language testing. For example, Gardiner and Howlett (2016) reported that of their participants,

"only the two Saudi students spoke in favor of typed responses and speaking onto a computer" (p.90). Some participants in the present study referred to their personal styles in relation to their testing mode preferences. For example, Aziz said that he is that type of personality who does not like dealing with the computer, as mentioned in the above quote. Also, Talal's (A/3/11) response in Example 40 goes in line with Aziz:

Example 40:

انا شخص اجتماعي وأحب اتكلم مع الناس ، حتى في الامتحانات ، ما أحب الاختبارات اللي عالكمبيوتر لان توترني لأن هذه هي شخصيتي وأسلوبي .أعتقد أن هذا يرتبط كثير بالشخصية ، بعض الأشخاص يحبون التعامل مع الأجهزة و الكمبيوتر ، وغيرهم زيي يحب يتكلم مع ناس حقيقين

I am a very social person and I like talking to real people, even in exams, I do not like computerized exams, they are very stressful for me because this is my personality and style. I think this very much related to personality, some people like dealing with machines and computers, others like me prefer talking to people.

Perceptions Towards VOCI

Participants' comments about the VOCI indicated a desire for more interaction.

Interaction

In terms of the test takers' perceptions towards VOCI, there is one sub-theme related to interaction (ambiguity of some questions and explanation). Odai (A/1/8) explained how some of the VOCI's questions were hard to understand, as shown in Example 41:

Example 41:

الفيديو كان ممتع لكن بعض الاسئله مافهمتها ماقدرت افهم ايش المقصود

The video is an interesting test, but for me I found it hard to understand some questions and I could not understand the intended meaning of the question.

While it is possible that some of the questions were unclear for some participants, this may be related to their listening comprehension skills, especially since not all participants claimed that some questions were unclear. However, it may also be related to the content of the questions. As claimed by Suryaningsih (2014) "the content, the manifestation, and the way tests are conducted are important matters to be reviewed carefully" (p.95). I agree with Suryaningsih that test content is important as it can stand as a barrier for test takers, as lack of content knowledge may prevent them from adequately displaying their proficiency. Talal (V/1/11) addressed the VOCI's lack of explanation in Example 42:

Example 42:

I needed more explanation for that question, and it was not possible to ask the computer for that.

Talal believes that one of the shortcomings of the VOCI is that if he does not understand the question, he cannot ask the computer to paraphrase the question. He went on in Example 43 to say:

Example 43:

This video test is not really good. I am not trying to be disrespectful, but the reason why I said this is sometimes I did not catch the questions, so I would not be able to communicate with the person who asked me for clarification.

Test structure

The theme of test structure had multiple sub-themes that specifically related to VOCI perceptions, including artificiality, formality, technology problems, contextualization, test score, and test goal. These sub-themes are illustrated in the examples below.

Example 44:

Artificiality was mentioned by Talal (V/2/11):

In the video, the voice is different due to the effect that it was recorded, it was intelligible, *yet it had that effect that made it look artificial*.

Louma (2004, p.168) emphasized that "the lack of reciprocity in tape-based testing can seem artificial to the examinees." While test developers could try to make semi-direct testing instruments more interactive, it might be very challenging to make those tools as interactive as face-to-face testing.

Formality was also mentioned by other participants. Alabad (V & A/2/13) thinks the VOCI is more formal than the OPI, as in Example 45:

Example 45:

The video test was unusual because you are talking to the computer, which means

you are talking to yourself, nobody is around, of course this makes you feel that this is a test not a natural conversation. But, in the video it looks more formal because it is similar to other computerized tests.

Alabad refers to the VOCI as "unusual," and "formal." He was asked through personal communication "why do you think talking to the computer is unusual and formal?" His response in Example 46 follows:

Example 46:

هو غير اعتيادي لان انت تتكلم مع الكمبيوتر هو صح ان حاليا ان نتكلم مع اجهزه زي مثلا سيري واليكسا لكن هذي البرامج او الاجهزه لها حدود معينه في المحادثه هي تكون مبرمجه لحد معين من الجمل لكن اذا تبغا محادثه من طرفين ماتقدر تسوي هذي المحادثات مع اجهزه عشان كذا انا اشوف ان غير اعتيادي بالنسبه ليش قلت ان الفيديو رسمي شوي لان الجهاز مايقدر يضحك او بتسم او يخلي جو المحادثه اقل رسميه انت مجرد تسمع السؤال وتجاوب فقط عكس المقابله كانت اللي تقابلني تبتسم وتعلق وهذا لطف الجو وخلاه اقل رسميه وقلل من التوتر

It is unusual because nobody makes conversation with a computer, it is true that we talk to the computers nowadays, I mean like talking to SIRI and Alexa, but it is only minimal amount of conversation, they programmed those machines to have limited capacities of sentences. But if you want reciprocal communication, you cannot do that with a machine, so that is why I see it as "unusual." It is also formal because you hear the question, then you answer, no further

communication, the machine cannot smile, or laugh, or say something that makes the situation less formal, unlike the interview, the interviewer was smiling, making comments, and replying back and that makes the atmosphere less stressful and less formal (Personal communication, October 19th, 2019).

Technology problems is another sub-theme related to test structure that was mentioned in connection with the VOCI. Aziz (A/2/7) actually experienced technology problems during the VOCI, as he explained in Example 47:

Example 47:

في بداية الفيديو الكمبيوتر علق فاضطرينا نعيد تشغيله ونبدا من جديد و هذا الشي فعلا حصل معي في اختبار حقيقي اكيد بتوتر

At the beginning of the video, the computer was freezing, and we had to restart it, then we started over again, it went well after that, but if I was actually taking a test and that happened to me, I will panic and get very nervous.

Although the OPI and VOCI in this study were used only for research purposes, some participants still focused on their scores when asked about their attitudes toward the tests, as evidenced by Salman's (V/2/8) response in the VOCI, as expressed in Example 48:

Example 48:

I am not sure which test is better in terms of scores because I need to see my results in each test then I can decide.

Despite some of the complaints about ambiguity, the fact that VOCI provided context for all of the questions made it easier for some of the participants to understand the questions. Nasser (V/2/12) appreciated this aspect of the VOCI in Example 49:

Example 49:

> *Actually for me, this test gave me examples before they ask me the question. Or not examples, maybe conversations, then they ask me the question, so that made me understand the question accurately, even though there are some words that I did not understand.*

Context is very important for test takers, because it helps them understand the questions and hence enables them to provide relevant answers. Students in Gardiner and Howlett's (2016) study commented about the difficulty in transitioning from one section of the test to another because there was a lack of context, "making idea development challenging" (pp.88-89).

Two participants reported that the topics in the VOCI are familiar, meaning that they had relevant knowledge that would help them answer the questions, as described by Mohammed (V/2/5) in Example 50:

Example 50:

> *This test is more about information and experiences; it is more about issues that we are very familiar with.*

Degree of difficulty was another way that participants addressed the sub-theme of topic familiarity. Odai (V/2/8) said he was thinking about other tests he has taken before the VOCI and in Example 51 he said that he considers the VOCI to be easier than those tests:

Example 51:

I think this test is easier.

Although Odai stated that the VOCI was an easier test, it was not clear what aspects of the test he considered easy. For that reason, Odai was asked to clarify his answer (Personal communication, October 20th, 2019), and responded in Example 52:

Example 52:

الاختبار سهل من ناحيه الاسئله مو كل الاسئله بس اكيد اغلبها. كمان الأساليب الموجوده في الفيديو كانت مفيده اقصد بالاساليب زي استخدام الصوره وكمان اغلب الاسئله قبلها في محادثه ونص يبين لك عن ايش السؤال. بالنسبه لي كان بس في سؤال واحد ما كان عندي اي معلومات عنه فهمته السؤال لكن ماعندي عنه اي افكره اللي هو السؤال عن فري تريد

The test is easy in terms of the level of the questions, not all of the questions but definitely the majority. Also, the strategies used in the video were helpful, and by strategies, I mean using pictures, and conversation that gives an idea of what the topic is about. For me, there was only one question that I did not have any knowledge about it, the free trade question.

Test goal is another sub-theme of test structure that Odai (V/2/8) mentioned in Example 53:

Example 53:

I think it is testing something else, other than skills, I do not know what exactly.

This is not surprising since the participants in this study did not take the VOCI to get admitted to their program or as part of an application for a job; they took it for the researcher to carry out her research, whose purpose was unknown to them. It appears that understanding test goals is important for test takers, as it might affect their performance. This finding is supported by Brown (2007), who stated that "to achieve peak performance on a test, a learner needs to be convinced that the test is indeed testing what it claims to test" (p. 449).

Critical thinking skills is another sub-theme related to test structure, in reference to the variety of question topics. In Example 54, Salman (V/2/8) reported:

Example 54:

This test is good because it asks me questions about different topics that made me do critical thinking in my brain.

Salman believes that the VOCI activated his critical thinking skills because of the variety of questions in this test. Perhaps he focused on critical thinking skills with regard to the VOCI because this test has a larger variety of questions than the OPI, where the interviewer sometimes asks a series of questions on the same topic in response to the interviewee's previous utterance.

Test takers' personal affective factors

The sub-themes that show test taker's personal affective factors include lack of motivation, boredom, and face image.

Ahmad (A/3/8) talked about lack of motivation during the VOCI in Example 55:

Example 55:

لما سويت الفيديو كنت بس مجرد اسجل اجاباتي ماكنت اتكلم بشكل طبيعي لان اعرف ان ماحد قاعد يسمعني هو عمليا كنت اقدر اتكلم واقول اي شي بدال ما اكون ساكت ولكن بصراحه ماكنت متحمس كثير للكلام ، ضحك

When I did the video, I was just recording my answers, not speaking naturally because I know nobody was listening to me. Technically I could have said anything just to talk instead of just being silent, but honestly, I was not motivated to do that, (laughter.)

This was not a high-stakes test, which probably contributed to the lack of motivation for some participants. Ahmad was asked about the reasons for his lack of motivation and his response appears below in Example 56:

Example 56:

ايه ما كان عندي حماس كثير لاني مافي سبب يخيك تتحمس تتكلم مع نفسك غير لما يكون قدامك شخص يسألك ويعلق على اجاباتك ويوريك انه مستمع بالكلام معاك هذي الاشياء تحمسك للكلام

Yes I did not have motivation because I do not see a reason or a motive to speak to myself, but when there is a person in front of you, that person will ask you

questions, give you comments, or show interest and all of these will make you motivated to talk.

Suryaningsih's (2014) participants also reported that they had no motivation when they took the IELTS (direct test), as they wanted only to try the test and they did not take it for real purposes. Similarly, my participants reported lack of motivation when they took the semi-direct VOCI, but interestingly they did state that they were motivated during the OPI. It is possible that they did not view the OPI as a test, as it was more like a conversation with a professor in her office.

Boredom is another subtheme listed under personal affective factors that was mentioned with respect to the VOCI. Alali (A/3/10) explained in Example 57 that the VOCI was boring for him:

Example 57:

عشان اكون صريح حسيت بملل في نص الاختبار وصلت مرحله اني بس ابغا اخلص

I have to be honest that I felt bored in the middle of the test, I reached a point that I just want it to finish.

Alali was asked why he thinks the VOCI was boring. In Example 58, he referred to its length and the lack of personal interaction:

Example 58:

هو ممل لان كان طويل وكنت اتكلم مع الكمبيوتر لو كنت اتكلم مع شخص اكيد بيكون ختلف ويمكن مااحس بالملل لان ما حاكون افكر بالوقت لكن لما اتكلم مع الكمبيوتر بدون شخص يرد علي شي ممل مو بس في الاختبارات في الحياه بشكل عام

It is boring because it was long, and I was talking to the computer, if I was talking to a person, that is different because I would not be bored, I will not think about time, but talking to a computer with no one responding is boring, not only in tests, but in life in general.

Another interesting sub-theme of the personal affective factors is the concept of face. Alabad (A/3/13) explained in Example 59 that he felt relaxed because there was no person to impress:

Example 59:

شي واحد في الفيديو عجبني اني مانحرج لما اتكلم يعني اذا قلت شي اهبل او مانطقت كلمه صحيحه او حتى ماعندي معلومات عن موضوع معين ما أهتم لان انا لحالي ما افكر

One thing in the video that I like is I do not feel embarrassed about what I say, I mean if I say something silly or mispronounce a word or do not have knowledge about the topic, I do not care because I am just by myself.

Alabad emphasizes that VOCI was more face-saving for him because he took the test alone. He considers the absence of a human interviewer as an advantage, as it is a way for him to avoid embarrassment and low self-esteem. This preference is mainly related to the personality of test takers. For example, perhaps a VOCI would be more appropriate for an introvert, whereas OPI might be more appealing to an extrovert. Amengual-Pizarro and García-Laborda (2012) found that "many test-takers described themselves as shy or introverted and pointed out they felt more relaxed before a computer without the presence of an examiner" (p.31). However, in this study, Alabad was the only participant who said he feels more comfortable in front of the computer, while the rest of

the participants reported that they were more comfortable during the face-to-face interviews.

Interestingly, participants further compared their experiences on the VOCI with TOEFL, another semi-direct test. For this reason, I include here a second level of comparisons between VOCI and TOEFL, even though the focus of this research is to compare perceptions toward VOCI and OPI. Comparison of these two semi-direct tests may point to factors affecting test preference that are unrelated to directness, such as topic familiarity and bias due to national origin, both of which were mentioned by participants while comparing the two semi-direct tests.

In analyzing the VOCI-TOEFL comparisons, I identified four subthemes relating to the theme of test structure (context, degree of difficulty, content, and topics). The following quotes illustrate their comparisons.

Degree of difficulty/ Decontextualization

Odai (V/2/8) began in Example 60 by referring to the VOCI:

Example 60:

In this test there are people talking together, then they ask you the question, in the TOEFL, they just ask the question, they do not give you conversation. It is not like the tests I have taken before, yeah, I think this test is easier than TOEFL.

Content/ Topics

Mohammed (V/2/5) also compared the VOCI with TOEFL in his response in Example 61:

Example 61:

TOEFL focuses on my skills, but this one is more about information and experiences, it is more about issues that we are very familiar with, but in TOEFL sometimes they asked me questions that I have never heard about before.

Salman reported that the VOCI is different from other tests he took before, and in Example 62 he explained why he liked it:

Example 62:

I took the TOEFL and in terms of speaking and listening, this test is good because it asks me questions about different topics that made me do critical thinking in my brain.

These findings about TOEFL and IELTS perceptions support those of Suryaningsih (2014). However, the present findings add more sub-themes to test takers' perceptions towards direct OPI and semi-direct VOCI. In Suryaningsih's study, participants viewed the IELTS as a more positive test than the TOEFL. In another study done by Soureshjani, Riahipour, and Safikhani (2012), technology involved in the semidirect test (TOEFL) was also addressed. They stated that "although technology-based test taking can be a great help in the more effective, practical, and efficient test taking, it can also be a disaster for some" (p. 25).

We can conclude from the findings that participants have different perceptions towards the direct test (OPI) and the semidirect test (VOCI). While the participants have

some positive perceptions towards the VOCI, they still reported that their preferred testing mode is the direct testing (OPI).

After learning about the test takers' perceptions towards OPI and VOCI, it would be interesting to see if there is any relation between their preferences and perceptions and their performance in terms of the CAF measures used in this study.

The results indicate that whereas the majority of the participants had more complex ASUs in the VOCI, they were less fluent and less accurate than in the OPI. The following graph illustrates the qualitative relation between CAF measures and testing mode.

PARTICIPANTS' PREFERENCES AND THEIR PERFORMANCE

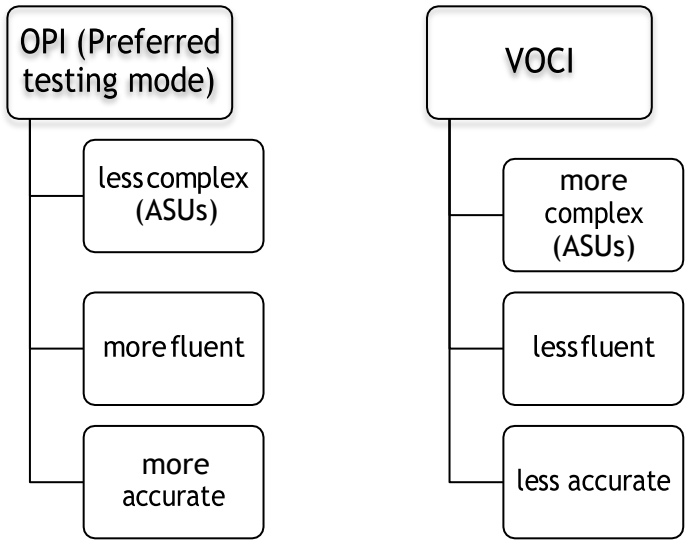

Figure 10. Qualitative relation between CAF measures and testing mode (OPI vs VOCI)

The previous figure shows that the participants stated that they prefer OPI, and they actually had higher accuracy and fluency in this mode. As for the complexity

measure, participants' responses on the VOCI had higher complexity than on the OPI, but the difference was not statistically significant. One could assume that the higher positive attitude towards the direct testing mode (OPI) could have led to higher accuracy and fluency in the OPI. While we need further investigation to claim that test preference could affect test performance, it also hard to disagree with this assumption. Participants mentioned that they felt more comfortable and more motivated talking to a person, which could make them less anxious and lead to better performance in terms of accuracy and fluency. I think the accuracy measure is of special interest here, because most non-native speakers try to avoid making errors, especially when talking to a native speaker. This issue could have made the participants more conscious about their production during the OPI, where they were being interviewed by a native speaker. In addition, participants mentioned that the presence of the interviewer made them more engaged in the discussion. This engagement likely meant that they thought less about the length of their responses, focusing instead of being intelligible to the interviewer, in terms of lexical choice and grammaticality.

CHAPTER VI

CONCLUSION

The current research aimed to determine whether direct (OPI) and semi-direct (VOCI) oral proficiency tests differ in terms of CAF specific measures. It further examined the relation between task type and CAF measures. In addition, the study investigated participants' preferences towards direct and semi-direct tests and examined the relationship between these preferences and test performance, as determined by specific CAF measures.

Nine Saudi participants, majoring in different engineering programs, were recruited for the study. Four instruments were used: OPI, VOCI, an online background survey, and Arabic interviews. Using an exploratory sequential mixed method approach, the researcher began with a qualitative phase that consisted in which the OPI and VOCI were administered. In the quantitative phase, the researcher then explored the perceptions and preferences of participants towards both testing modes. Subsequently, parts of the participants' responses in the OPI and VOCI were coded for complexity, accuracy and fluency. These CAF measures were then analyzed using statistical and inferential statistics, to determine whether differences between the testing modes were significant, and to examine the relation between task type and CAF measures in both testing modes.

Findings revealed a significant difference in the accuracy (EF-ASU) and fluency measures (SP) between the OPI and VOCI. The significant differences in participants'

performance (EF-ASU, and SP) in two different testing modes support the findings of some previous studies, including Brooks and Swain (2013, 2015), O'Loughlin (1995), Shohamy (1994) and Ure (1971). However, the complexity measure did not show any significant differences, despite noticeably higher VOCI scores for ML-ASU and MS-TTR. Therefore, it is possible that both tests (VOCI and OPI) are testing the same aspect of grammatical complexity and hence they can be used interchangeably.

On the OPI, ASUs had higher accuracy in terms of the proportion of EF-ASU, and higher fluency in terms of the number of silent pauses. It is possible that both OPI and VOCI could be equivalent in terms of measuring the complexity of oral proficiency. This makes the VOCI a potentially effective testing instrument in contexts where the availability of a certified language tester is not practical. For example, it could be used in Saudi Arabia for oral proficiency testing, especially since participants viewed the VOCI positively, despite their preference for communicating with a human tester.

Findings revealed that the narrative task impacted the MS-TTR in the VOCI and the number of the SP in the OPI. Since the narrative task showed a significant effect on both the complexity and fluency measures, test developers should include a variety of task types in any oral proficiency test. Tests' raters should also make sure to rate testees' responses using different tasks, in order to have a better picture of the testees' oral proficiency performance. The significant effect of task type on CAF measures supports the approaches followed by TOEFL iBT and IELTS in using different tasks and different language functions. Bachman and Palmer (1996, p.10) stated

> If we want to use the scores from a language test to make inferences about individuals' language ability, and possibly to make various types of decisions, we must be able to demonstrate how performance on that language test is related to language use in specific situations other than the language test itself...That is, we need a framework that enables us to use the same characteristics to describe what we believe are the critical features of both language test performance and non-test language use.

Study participants expressed a preference for the direct testing mode (OPI). Previous studies (Jeong, 2003; Kamal et al., 2012; McNamara, 1987; Qian, 2009; Shohamy, Donitsa-Schmidt, & Waizer, 1993; Stansfield, Kenyon, Paiva, Doyle, Ulsh, & Cowles, 1990; Suryaningsih, 2014) have also found that test takers prefer direct testing and attributed this preference to direct tests being more communicative and interactive, as they involve human interaction. However, this study illuminates additional sub-themes that may influence perceptions towards both testing modes. Participants did not prefer the VOCI, despite perceiving some of its aspects positively. For example, participants claimed that VOCI is face-saving because if they make a mistake or mispronounce a word, they will not feel embarrassed because there is no person in front of them.

Kenyon and Tschirner (2000) stated that one disadvantage of the direct OPI is that it requires a "highly skilled and thoroughly trained individual" (p.99) to administer, whereas semi-direct tests require training only for test raters. This disadvantage is especially relevant in many EFL contexts, where access to certified language testers is

limited. For example, in Saudi Arabia, where my participants are from, testing oral proficiency using the VOCI would be more practical than OPI, which requires a certified language tester. Given the fact that the participants prefer the OPI, I still recommend using the VOCI because of its practicality in the Saudi context as it can be administered without a certified rater. Most universities have a preparatory year program that requires taking an intensive English course; however, oral proficiency is never tested in many universities, most likely because few Saudi institutions have access to certified language testers, which can be expensive. Having a language tester do individual interviews is also extremely time-consuming, and difficult to coordinate considering the large Saudi class sizes. In comparison, VOCI administration is much simpler to coordinate, as it can be administered to large groups of test takers simultaneously.

As I mentioned earlier in this dissertation, testing oral proficiency in the Saudi context seems to be neglected. Alharbi and Surur (2019) stated "attempts to evaluate assessment techniques and procedures, especially for speaking skills, are lacking in a Saudi context" (p.1). There are some factors that make testing oral proficiency in Saudi context a challenging task; one of those challenges is the unavailability of a reliable testing instrument. Sharma (2016) also reported that the evaluation system in Saudi Arabia does not include the speaking skill.

Given the context of testing oral proficiency in Saudi Arabia, and the fact that grammatical complexity did not show any differences between the VOCI and OPI, I believe that the VOCI will serve as a practical testing tool for several reasons. First, VOCI is a ready-made test to be used in any classroom. Second, it is based on the

ACTFL proficiency guidelines, which makes it a reliable measure of oral proficiency. In addition, VOCI can be administered to any class size at the same time. Furthermore, teachers can choose appropriate questions of the VOCI that match the students' proficiency levels, instead of giving the whole test.

Based on the findings of this study, complexity measures did not show any statistically significant differences between the OPI and VOCI. This means that both tests are equivalent in terms of the syntactic and lexical complexity. In addition, several participants reported some positive aspects of the VOCI. For example, one of the participants stated that the VOCI enhances the critical thinking skills. Another participant claimed that VOCI is face-saving as he will not be embarrassed if he makes a mistake or does not know how to answer the question. These findings make the VOCI a possible testing instrument of oral proficiency in the Saudi context.

VOCI can also be used in a smaller scale, inside the classrooms, to conduct needs analysis. Teachers can use the VOCI to examine the oral proficiency skill of their students and set their course objectives accordingly. This way, teachers could have a clear picture of the individual differences among students in terms of their oral proficiency levels so that they could address the challenges of their students. Students will also be interested to get feedback about their oral proficiency performance.

At universities where there are no OPI raters, teachers could be trained to be VOCI raters. And in universities where no OPI testers could administer an OPI, the VOCI can be administered, since the test is all ready to be administered (no training is necessary). Kenyon and Tschirner (2000) suggested that "it is easier to learn to rate the

SOPI at least somewhat reliably within a relatively short period of time than it is to learn to administer and rate the OPI" (p.99). The same can be assumed for the VOCI.

The current study also raises some questions, such as whether the directness of the tests is what shaped the participants' perceptions towards the tests, or whether other factors affected those perceptions, such as gender, age, or even national origin. Since the participants reported lack of motivation during the VOCI, future studies could investigate whether motivation would increase if the test were given in a context with higher stakes, such as a course grade or academic program admission.

Hill (1998) claimed that test takers' feedback can be an indicator of test validity and acceptability. She also emphasized that test takers' feedback can support statistical analysis by helping to identify problematic issues with the test. Based on Hill's argument, we can say that the feedback received from the Saudi participants regarding the OPI and VOCI tests did really inform us that while Saudi participants preferred the OPI, they also have some positive perceptions towards the VOCI, which indicates test acceptability as stated by Hill (1998). The findings revealed that participants had both positive and negative perceptions toward those tests, which can help raise awareness for test developers and administrations to consider when designing direct and semi-direct tests. There

All of the participants had positive perceptions of the OPI, with several reporting that it resembled real communication in daily life. This study corroborates the findings of previous studies that reported a preference for direct testing (Jeong, 2003; McNamara, 1987; Qian, 2009; Shohamy et al., 1993; Stansfield et al., 1990). However, it is one of the

few studies that used oral proficiency measures to analyze test performance. Qian (2009) used a questionnaire that asked his participants if they would perform better in the test that they prefer. However, this is not really testing their performance, it is basically collecting their views or perceptions, or expectations regarding the relation between testing performance and testing mode. This study adds to the literature that participants in this study did not show significant differences in the complexity measures. This encourages future researchers to investigate if the test takers' preference towards oral proficiency testing mode would affect their testing performance in terms of complexity using other types of direct and semi-direct testing modes. Future studies could also compare test takers' preferences with test takers' performance in terms of the other measures, such as accuracy and fluency.

Study participants claimed that they preferred OPI, and analysis of their production showed that they had more accurate and fluent ASUs in the OPI. This finding partially supports the findings of Qian (2009) who found that if test takers' state of mind is negatively affected by the testing mode, their affective filter may interfere with their test performance.

Limitations of this study include sample size. Future studies are encouraged to have a larger sample size. Also, all of the participants were Arabic speakers from Saudi Arabia. Future studies could include and possibly compare the performance of speakers of other languages.

In this study, I used the number and mean length of ASUs, MS-TTR, and EF-ASUs for accuracy, and SP and FP for fluency. I encourage future studies to use the same

CAF measures with different populations, and it would be interesting to compare their findings with the findings of this study.

REFERENCES

Ahmadian, M. J., & Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research, 15*, 35-59.

Ahmad, A. (2014). Kumaravadivelu's framework as a basis for improving English language teaching in Saudi Arabia: Opportunities and challenges. *English Language Teaching, 7*, 96-110.

Ahmed, S., & Alamin, A. (2014). Assessing speaking ability in academic context for fourth year Taif university students. *International Journal of English Linguistics, 4,* 97-103.

Al Asmari, A. (2013) Saudi university undergraduates' languagelLearning attitudes: A preparatory year perspective. *International Journal of Asian Social Science,3,* 2288-2306.

Alasmari, A., & Khan, Sh. (2014). World Englishes in the EFL Teaching in Saudi Arabia. *Arab World English Journal, 5*, 316-325.

Alderson, J. C. (1988). New procedures for validating proficiency tests of ESP? Theory and practice. *Language Testing, 5*, 220–32.

Alderson, J. C., & Banerjee, J. (2002). State of the art review: Language testing and assessment (Part 2). *Language Teaching*, *35*, 79–113.

Alderson, C. (2004). The shape of things to come: Will it be the normal distribution? In M. Milanovic & C. J. Weir (Eds.), *Studies in language testing 18: European language testing in a global context* (pp. 1–26). Cambridge, UK: Cambridge University Press.

Alfallaj, F., & Al-Ahdal, A. (2017). Authentic Assessment: Evaluating the Saudi EFL Tertiary Examination System. *Theory and Practice in Language Studies, 7*, 597-607.Al Hajailan, D. T. (2003). *Teaching English in Saudi Arabia*. Riyadh: Aldar Al sawlatiah.

Alharbi, A., & Surur, R. (2019). The effectiveness of oral assessment techniques used in EFL classrooms in Saudi Arabia from students and teachers point of view. English Language teaching, 12,1.

Alhmadi, N. (2014). English speaking learning barriers in Saudi Arabia: A case study of Tibah University. *Arab World English Journal, 5*, 38-53.

Al Hassan, N. (2015). Saudi EFL university instructors' barriers to teaching the speaking skills: Causes and solutions. Unpublished M.A. Thesis, Al-Imam Muhammad Ibn Saud Islamic University

Al-Jarf, R. (2006). Large student enrollments in EFL programs: Challenges and consequences. *Asian EFL Journal Quarterly, 8*, 8-34.

Al-Ma'shy, A.A. (2011). Causes of EFL speaking weakness in Saudi secondary schools in Al-Gunfuthah City. Unpublished M.A. Thesis, King Saud University.

Al Mineeai, A. (2013) Problems of speaking faced by female EFL students in Bisha secondary schools from teachers' and students' perspectives. Unpublished M.A. Thesis, Al- Imam Muhammad Ibn Saud Islamic University.

Al-Nasser, A.S. (2015). Problems of English language acquisition in Saudi Arabia: an exploratory-cum-remedial study. *Theory and Practice in Language Studies, 5*, 1612-1619.

Al-Nofaile, H. (2010). The Attitude of teachers and students towards using Arabic in EFL classrooms in Saudi public schools. *Novitas Royal Research on Youth and Language, 4*, 64-95.

Al-Seghayer, K. (2011). *English teaching in Saudi Arabia: Status, issues, and challenges*. Riyadh, Saudi Arabia: Hala Printed Co.

Al-Shumaimeri, Y. A. N. (2003). A study of classroom exposure to oral pedagogic tasks in relation to the motivation and performance of Saudi secondary learners of English in a context of potential curriculum reform. Unpublished Ph.D. thesis, submitted to University of Leeds, United Kingdom.

Alsudais, A. (2017). Teaching English as a foreign language: The case of Saudi Arabia. *European Journal of English Language and Literature studies,5*, 18-27.

American Council on the Teaching of Foreign Languages. (2017, November 14). Retrieved from https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012/english/speaking#advanced

Amengual-Pizarro, M., & García-Laborda, J. (2017). Analysing test-takers' views on a computer-based speaking test. *Profile Issues in Teachers' Professional Development, 19*, 23-38.

Arnett, K. & Haglund, J. (2001). American Council on the Teaching of Foreign Languages Oral Proficiency Interview. *The Canadian Modern Language Review, 58*, 312-318.

Ata, A. (2015). Knowledge, education, and attitudes of international students to IELTS: A case of Australia. *Journal of International Students, 5*, 488-500.

Bacha, N.N. (2002). Developing Learners' academic writing skills in higher education: A study for educational reform. *Language &Education, 16*, 161-177.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (1998). Appendix: Language testing -SLA research interfaces. In L.F. Bachman and A.D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 177-95). Cambridge: Cambridge University Press.

Bachman, L. F., & Cohen, A. (1998). *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Savignon, S. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal, 70*, 380-390.

Bahanshal, D. (2013). The effect of large classes on English teaching and learning in Saudi secondary schools. *English Language Teaching, 6*, 49-59.

Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT and real-life academic speaking activities. *Language Assessment Quarterly, 11*, 353–373.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly, 25*, 587–603.

Brown, A. (1993). The role of test-taker feedback in the development process: Test takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing, 10,* 277–304.

Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy* (3rd ed.). New York: Pearson Longman.

Brown, A., Cox, T., & Thompson, G. (2017). A comparative discourse analysis of Spanish past narrations from the ACTFL OPI and OPIc. *Foreign Language Annals, 50*, 793-807.

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

Bygate, M. (1999). Quality of language and purpose of task: Patterns of learners' language on two oral communication tasks. *Language Teaching Research, 3*, 185-214.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Task-based learning: Language teaching, learning and assessment.* (pp.23-48). London; Longman.

Chambers, F. (1997). What do we mean by fluency? *System, 25*, 535-544.

Chapelle, C. A. (2003). *English language learning and technology*. Amsterdam: John Benjamins.

Chapelle, C., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

Chau, H. P. (2014). *The effects of planning with writing on the fluency, complexity, and accuracy of L2 oral narratives* (Publication No. 3645908) [Doctoral dissertation, Michigan State University]. ProQuest Dissertation Publishing.

Clark, J. L. D., & Li, Y. C. (1986). *Development, validation, and dissemination of a proficiency-based test of speaking ability in Chinese and an associated assessment model for other less commonly taught languages.* Washington, DC:

Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED 278 264.)

Clark, J. L. D., & Clifford, R. (1988). The FSI/ILR/ ACTFL proficiency scales and testing techniques: Development, current status, and needed research. *Studies in Second Language Acquisition, 10,* 129-147.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Creswell, J. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Thousand Oaks, California: SAGE Publications.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, *11,* 367-222.

Crookes, G. (1990). The utterance, and other base units for second language discourse analysis. *Applied Linguistics*, *11*, 183-199.

Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals, 23*, 11-22.

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

Duff, P. (1986). Another look at interlanguage talk: taking task to ‘task'. In R.R. Day

(Ed.), *Talking to Learn: Conversation in Second Language Acquisition* Rowley, MA: Newbury house. 237-326.

Educational Testing Services (ETS). (2019, October 19th). Retrieved from https://www.ets.org/toefl/ibt/about/content/

Ehret, K., & Szmrecsanyi, B. (2016). An information-theoretic approach to assess linguistic complexity. In R. Baechler and G. Seiler. (Eds.) *Complexity, isolation and variation*(71-94). Berlin/Boston, MA: De Gruyter.

Ejzenberg, R. (1997). The role of task structure on oral fluency assessment. Presented at the 28th Annual Meeting of TESOL USA, Baltimore: ERIC Document Reproduction Service No. ED 334 238.

Ejzenberg, R. (2000) The juggling act of oral fluency: A psycho- sociolinguistic metaphor. In H. Riggenbach. (Ed.), *Perspectives on fluency* (287-313), Michigan: The University of Michigan Press.

Ellis, R. (1990a). Instructed language learning. Oxford: Blackwell.

Ellis, R. (1990b). Individual styles in classroom second language development. In J. De Jong and D. Stevenson (Eds.), *Individualizing the assessment of language abilities*. Clevedon, Avon: Multilingual Matters.

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

Ellis, R. (2005). *Planning and task performance in a second language* (Language learning and language reaching; v. 11). Amsterdam; Philadelphia: John Benjamins Pub.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, *30*(4), 474–509.

Ellis, R., & Barkhuizen, G. (2005). *Analyzing learner language*. New York, NY: Oxford University Press.

Ellis, R., & Yuan, F. (2005). The effects of careful within-task planning on oral and written task performance. In R. Ellis (Ed.), Planning and task performance in a second language. John Benjamins.

Fareh, S. (2010). Challenges of teaching English in the Arab world: Why can't EFL programs deliver as expected? *Procedia - Social and Behavioral Sciences, 2*, 3600-3604.

Farooqui, S. (2007). Developing speaking skills of adult learners in private universities in Bangladesh: Problems and solutions, *Australian Journal of Adult Learning,47*, 94-110.

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

Freed, B. F., Dewey, D. P., Segalowitz, N., & Halter, R. (2004). The language contact profile. *Studies in Second Language Acquisition, 26,* 349-356.

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

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*, 354-375.

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

Fulcher, G. (1996). Testing tasks: issues in task design and group oral. *Language Testing, 13*, 23- 49.

Gaies, S. J. (1980). T-unit analysis in second language research: Applications, problems and limitations. *TESOL Quarterly, 14*, 53-60.

Gardiner, J., & Howlett, S. (2016). Student perceptions of four university gateway tests. *University of Sydney Papers in TESOL, 11,* 67–96.

Giles, W. (2016). *A comparison of fluency and complexity in two different kinds of oral test.* Paper presented at Twenty-fifth International Symposium on English Teaching, Taipei, Taiwan.

Gubaily, M. (2012). Challenges of teaching and learning spoken English in Yemen. *International Journal of Social Science Tomorrow, 1,* 1-8.

Guillot, M. (1999). *Fluency and its teaching*. Clevedon: UK: Multilingual Matters.

Halleck, G. B. (1992). The oral proficiency interview: Discrete point test or a measure of communicative language ability? *Foreign Language Annals, 25*, 227-231.

Halleck, G. B. (1995). Assessing oral proficiency: A comparison of holistic and objective measures. *The Modern Language Journal, 29*, 223-234.

Halleck, G. B., & Young, R. (1995). Video Oral Communication Instrument. Language Acquisition Resource Center (LARC).San Diego State University.

Halleck, G. (1996). Assessing Oral Proficiency: A comparison of holistic and objective Measures. *Modern Language Journal, 79*, 223-34.

Halleck, G. (2005). Unsubstantiated claims about the Oral Proficiency Interview. *Language Assessment Quarterly, 2*, 315-319.

Hamad, M. (2013). Factors negatively affect speaking skills at Saudi colleges for girls in the south. *English Language Teaching, 6*, 87.

Hammadou Sullivan, J. (2011). Taking charge: Teacher candidates' preparation for the Oral Proficiency Interview. *Foreign Language Annals, 44*, 241–257.

Hammerly, H. (1991). *Fluency and accuracy: Toward balance in language teaching and learning*. Clevedon, UK: Multilingual Matters.

Higgs, T., & American Council on the Teaching of Foreign Languages. (1984). *Teaching for proficiency: The organizing principle* (ACTFL foreign language education series). Lincolnwood, Ill.: National Textbook.

Hill, K. (1998). The effect of test taker characteristics on reactions to and performance on an oral English proficiency test. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 209-229). Mahwah, NJ: Lawrence Erlbaum.

Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *The Language Learning Journal, 36*, 153-166.

Hosseini, A. S., & Azarnoosh, M. (2014). Iranian EFL instructor's oral assessment practices: Purposes, methods, procedures. *Procedia-Social and Behavioral Sciences, 98*, 653-658.

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

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

Hu, X. (2018). Effects of task type, task type repetition, and performance criteria on L2 oral production. In M. Bygate (Ed.), *Learning language through task repetition* (pp. 143-169). Amsterdam: John Benjamins.

Huei-Chun, T. (2007). A study of task type for L2 speaking assessment. Paper presented at the Annual Meeting of the International Society for Language Studies (ISLS), Honolulu, HI.

International English Language System (IELTS). (2019, October 19th). Retrieved from

https://www.ieltsessentials.com/prepare/free-practice-tests/speaking

Isbell, D., & Winke, P. (2019). ACTFL Oral Proficiency Interview-computer (OPIc). *Language Testing, 36*, 467-477.

Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language Assessment Quarterly, 3,* 151-169.

James, G. (1988). Development of an oral proficiency component in a test of English for academic purposes. In A. Hughes (Ed.), *Testing English for university study (ELT Documents 127)* (pp. 111–133). Oxford, UK: Modern English Publications and

Javid, C. Z. (2011). Saudi medical undergraduates' perceptions of their preferred learning styles and evaluation techniques. *Arab World English Journal, 2*, 40-70.

Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *Modern Language Journal, 87*, 90-107.

Jeong, T. (2003). *Assessing and interpreting students' English oral proficiency using d - VOCI in an EFL context* (Publication No. 3088860) [Doctoral dissertation, The Ohio State University]. ProQuest Dissertation Publishing.

Jeong, H., Hashizume, H., Sugiura, M., Sassa, Y., Yokoyama, S., Shiozaki, S., & Kawashima, R. (2011). Testing second language oral proficiency in direct and semi-direct settings: A social-cognitive neuroscience perspective. *Language Learning, 61*, 675–699.

Johnson, M. (2000). Interaction in the Oral Proficiency Interview. Pragmatics. *Quarterly Publication of the International Pragmatics Association (IPrA), 10*, 215–231.

Johnson, M., & Tyler, A. (1998). Reanalyzing the OPI: How much does it look like natural conversation? In R. Young & A. Weiyun He (Eds.), *Talking and testing: discourse approaches to the assessment of oral proficiency* (pp. 27–51). Amsterdam: John Benjamins.

Kachru, B. B. (1992). *The other tongue: English across cultures(Ed.)*. Urbana: University of Illinois Press.

Kanga, K. (2012). Individual and paired oral proficiency testing: A study of learners' preference. Retrieved from https://d.lib.msu.edu/etd/973

Kenyon, D. M. (1998). An investigation of the validity of task demands on performance-based tests of oral proficiency. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 19-40). Mahwah, NJ: Lawrence Erlbaum.

Kenyon, D. M., & Stansfield, C. W. (1993). A method for improving tasks on performance-based assessments through field testing. In A. Huhta, K. Sajavaara, & S. Takala (Eds.), *Language testing: New openings* (pp. 90-102). Jyväskylä, Finland: University of Jyväskylä.

Kenyon, D., & Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. *The Modern Language Journal,84,* 85-101.

Khazaei, Z., Zadeh, A., &Ketabi, S. (2012). Willingness to communicate in Iranian EFL learners: The effect of class size. *English Language Teaching, 5*, 181-187.

Kissau, S. (2014). The impact of the Oral Proficiency Interview on one foreign language teacher education program. *Foreign Language Annals*, *47*, 527–545.

Kitajima, R. (2009). Negotiation of meaning as a tool for evaluating conversational skills in the OPI. *Linguistics and Education, 20*, 145–171.

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35*, 477–499.

Koike, D. A. (1998). What happens when there's no one to talk to? Spanish foreign language discourse in Simulated Oral Proficiency Interview. In R. Young & A. Weiyun He (Eds.), *Talking and testing*: *Discourse approaches to the assessment of oral proficiency* (pp.69-100). Amsterdam: John Benjamins.

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

Kortmann, B., & Szmrecsanyi, B. (2012). *Linguistic complexity: Second language acquisition, indigenization, contact*. Berlin/Boston, MA: De Gruyter.

Kowal, S., Wiese, R., & O'Connell, D. C. (1983). The use of time in storytelling. *Language and Speech*, *26*, 377–392.

Kuiken, F., Vedder, I., & Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. In I. Bartining, M. Martin & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLS and language testing research* (pp. 81-100). Amsterdam: European Second Language Association (EUROSLA).

Kuo, J., & Jiang, X. (1997). Assessing the assessments: The OPI and the SOPI. *Foreign Language Annals, 30*, 503-512.

Labov, W. (1997). Some further steps in narrative analysis. *The Journal of Narrative and Life History*, 7, 207–215. Retrieved from https://www.ling.upenn.edu/~wlabov/sfs.html

Language Testing International (LTI). (2019, March 10th). ACTFL speaking assessment: The Oral Proficiency Interview (OPI). Retrieved from https://www.languagetesting.com/oral-proficiency-interview-opi

Lantolf, J. P., & Frawley, W. (1985). Oral proficiency testing: A critical analysis. *Modern Language Journal, 69*, 337-3.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*, 590-619.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*, 579-589.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge, UK: Cambridge University Press.

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

Liton, H. (2012). Developing EFL teaching and learning practices in Saudi colleges: A review. *International Journal of Instruction*, *5*, 129-152.

Liskin-Gasparro, J. (1984a). The ACTFL proficiency guidelines: a historical perspective. In T. Higgs (Ed.), *Teaching for proficiency, the organizing principle* (pp. 11-42). Lincolnwood, IL: National Textbook Company.

Liskin-Gasparro, J. (1984b). The ACTFL proficiency guidelines: gateway to testing and curriculum. *Foreign Language Annals*, 17, 475-489.

Liskin-Gasparro, J. E. (2003). The ACTFL proficiency guidelines and the oral proficiency interview: A brief history and analysis of their survival. *Foreign Language Annals, 36*, 483-490.

Lowe, P. (1986). Proficiency: Panacea, framework, process? A reply to Kramsch, Schulz, and, particularly, to Bachman and Savignon. *Modern Language Journal, 70*, 391-397.

Lowenberg, P. H. (2002). Assessing English proficiency in the expanding circle. *World Englishes, 21*, 431-435.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing, 22*, 415–437

Luoma, S. (1997). *Comparability of a tape-mediated and a face-to-face test of speaking: A triangulation study* [Unpublished licentiate thesis]. Centre for Applied Language Studies, University of Jyvaskyla.

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

Malabonga, V., Kenyon, D. M., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing, 22*, 59–92.

Malone, M. E., & Montee, M. J. (2010). Oral proficiency assessment: Current approaches and applications for post-secondary foreign language programs. *Language and Linguistics Compass, 4,* 972-986.

McNamara, T. F. (1987). *Assessing the language proficiency of health professionals: Recommendations for the reform of the Occupational English Test (Report submitted to the Council of Overseas Professional Qualifications)*. Department of Russian and Language Studies, University of Melbourne, Melbourne, Australia.

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

Michel, M. (2011). Effects of task complexity and interaction on L2 performance. In P. Robinson (Ed.), *Second language task complexity: researching the cognition hypothesis of language learning and performance* (pp. 141–173). Amsterdam: John Benjamins.

Mirshahidi, S. (2017). *Predicting international teaching assistants' performance in a domain-specific test: The case of complexity, accuracy, fluency, and compensatory strategies* Publication No.10275408) [Doctoral dissertation, Oklahoma State University]. ProQuest Dissertation Publishing.

Moder, C. L. & Halleck, G. B. (1998). Framing the LPI as a speech even. In R. Young & A. W. He (Eds.) *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 117-146). Amsterdam: John Benjamins.

Noor, N. B. M., Muniandy, M. K., Shanmugan, S. K. K., & Mathai, E. J. (2010). Upper primary teachers' perceptions of PSLE English oral assessment. *English Language Teaching, 3,* 142-151.

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

Noubandegani, P. A. (2012). Students' perceptions of computerized TOEFL test. *Language Testing in Asia, 2*, 73–101.

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*, 557-581.

Okada, Y. (2010). Role-play in oral proficiency interviews: Interactive footing and interactional competencies. *Journal of Pragmatics, 42*, 1647-1668.

O'Loughlin, K. J. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing, 12*, 217–37.

O'Loughlin, K. J. (1997). *Direct and semi-direct tests of spoken language* [Unpublished doctoral thesis]. University of Melbourne.

O'Loughlin, K. J. (2001). *The equivalence of direct and semi-direct speaking tests*. Cambridge, UK: Cambridge University Press.

Omaggio, A. C., (1986). *Teaching language in context*: *Proficiency-oriented instruction*. Boston, Mass.: Heinle & Heinle.

O'Sullivan, B. (2012). A brief history of language testing. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 9–19), Cambridge, UK: Cambridge University press.

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

Pienemann, M., Johnston, M., & Brindley. G. (1988). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition,10*, 217-243.

Purpura, J. E. (2013). Assessing grammar. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 100-125). Malden, MA: Wiley.

Qian, D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly, 6*, 113-125.

Rababah, G. (2003). Communication problems facing arab learners of English: A personal perspective. *TEFL Web Journal,2*, 15-30.

Rahman, M., & Alhaisoni, E. (2013). Teaching English in Saudi Arabia: Prospects and challenges. *Academic Research International Journal, 4*, 112-118.

Raish, M. (2017). *The measurements of the complexity, accuracy, and fluency, of written Arabic* (Publication No. 10272590) [Master's thesis, Georgetown University]. ProQuest LLC

Rasti, I. (2009). Iranian candidates' attitudes towards IELTS. *Asian EFT Journal, 11*, 5.

Read, J. (2000). *Assessing vocabulary* (Cambridge language assessment series). Cambridge; New York: Cambridge University Press.

Reed, D. J., & Halleck, G. B. (1997). Probing above the ceiling in oral interviews: What's up there? In A. Huhta, V. Kohonen, L Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 225-38). Jyväskylä, Finland: University of Jyväskylä and University of Tampere.

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

Riggenbach, H. (1989). Evaluating learner interactional skills: Conversation at the micro level. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp.53-67). Amsterdam: John Benjamins.

Robinson, P. (1995). Attention, memory and the 'noticing' hypothesis. *Language Learning, 45*, 283-331.

Robinson, P. (2001a). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287-318). Cambridge, UK: Cambridge University.

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

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics, 45*, 237–57.

Robinson, P., Ting, S., & Urwin, J.J. (1995). Investigating second language task complexity. *RELC Journal, 26*, 62-79.

Robinson, P., Cadierno, T., & Shirai, Y. (2009). Time and motion: Measuring the effects of the conceptual demands of tasks on second language speech production. *Applied Linguistics, 30*, 533–554.

Robinson, P., & Gilabert, R. G. (2007). Task complexity, the Cognition Hypothesis and second language learning and performance. *International Review of Applied Linguistics in Language Teaching, 45*, 161-176.

Ross, S. & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*, 159–76.

Ross, S., & Kasper, G. (1998). Topic management in oral proficiency interviews. Paper presented at AAAL conference, Seattle.

Ruiz-Funes, M. (2014). Task complexity and linguistic performance in advanced college-level foreign language writing. *Task-Based Language Learning–Insights from and for L2 Writing, 7*, 163.

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

Sharma, V. (2016). How do productive skills of Saudi students affect EFL learning and teaching*? Asian Journal of Humanities and Social Sciences* 3, 91-99.

Shohamy, E. (2004). Assessment in multicultural societies: Applying democratic principles and practices to language testing. In B. Norton, & K. Toohey (Eds.), *Critical pedagogies and language learning* (pp. 72–92). Cambridge, Cambridgeshire: Cambridge University.

Shohamy, E., Donitsa-Schmidt, S., & Waizer, R. (1993). *The effect of the elicitation mode on the language samples obtained in oral tests*. Paper presented at the 15th Language Testing Research Colloquium, Cambridge, UK.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*, 99–123.

Shohamy, E., Gordon, C., Kenyon, D. M., & Stansfield, C. W. (1989). The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. *Bulletin of Higher Hebrew Education, 4*, 4-9.

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

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogical tasks, second language learning, teaching and testing* (pp. 167-185). New York: Pearson Education Limited.

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

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

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*, 185-211.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language learning, 49*, 93-120.

Soureshjani, K., Riahipour, H., & Safikhani, P. (2012). An investigation into the Iranian EFL language learners' attitudes on TOEFL iBT. *Language Testing in Asia, 2*, 1-15.

Stansfield, C. W. (1991). A comparative analysis of simulated and direct oral proficiency interviews. In S. Anivan (Ed.), *Current developments in language testing* (pp. 199–209). Singapore: RELC.

Stansfield, C. W. (1992). *ACTFL Speaking Proficiency Guidelines. ERIC Digest*. (ERIC Document Reproduction Service No. ED347852).

Stansfield, C. W. (1996). *SOPI test development handbook*. Washington, DC: Center for Applied Linguistics.

Stansfield, C. W., & Kenyon, D. M. (1991). *Development of the Texas oral proficiency test (TOPT). Final Report.* Washington, DC: Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED 332 522).

Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20,* 347–364.

Stansfield, C. W., Kenyon, D. M., Paiva, R., Doyle, F., Ulsh, I., & Cowles, M. A. (1990). The development and validation of the Portuguese Speaking Test. *Hispania, 73*, 641-651.

Surface, E., & Dierdorff, E. (2003). Reliability and the ACTFL Oral Proficiency Interview: Reporting indices of interrater consistency and agreement for 19 languages. *Foreign Language Annals, 36*, 507–519.

Suryaningsih, H. (2014). *Students' perceptions of International English Language Testing System (IELTS) and Test of English as a Foreign Language (TOEFL) tests*. Retrieved from https://pdfs.semanticscholar.org/7b8d/05a682a6fdeb7f55d20d0a144354dac21948.

Swain, M., Huang, L. S., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). *The speaking section of the TOEFL iBT (SSTiBT): Test-takers' reported strategic behaviors (TOEFL iBT Research Rep. No. 10).* Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RR-09-30.pdf

Tahaineh, Y.S. (2010). Arab EFL university students' errors in the use of prepositions. *MJAL, 2*, 76-112.

Tarone, E. (1985). Variability in Interlanguage use: A study of style-shifting in morphology and syntax. *Language Learning*, *35*, 373–404.

Tarone, E. & Parrish, B. (1988). Task-related variation in interlanguage: The case of articles. *Language Learning, 38*, 1-44.

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*, 133-150.

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

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

Test of English as Foreign Language (TOEFL). (2019, October 19th) Retrieved from https://www.ets.org/toefl/ibt/about/content/

Thompson, I. (1995). A study of interrater reliability of the ACTFL Oral Proficiency Interview in five European languages: data from ESL, French, German, Russian and Spanish. *Foreign Language Annals, 28*, 407–422.

Thompson, I. (1996). Some misconceptions about communicative language teaching. *ELT Journal*, 50, 9-15.

Tonkyn, A. (2012). Measuring and perceiving changes in oral complexity, accuracy and fluency: Examining instructed learners' short-term gains. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 221-244). Amsterdam: John Benjamins.

Valdman, A. (1988). Introduction. *Studies in Second Language Acquisition, 10*, 121-128.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral Proficiency Interviews as conversation. *TESOL Quarterly, 23*, 489-508.

Vercellotti, M. L. (2012). *Complexity, accuracy, and fluency as properties of language performance: The development of the multiple subsystems over time and in relation to each other* (Publication No. 3529566) [Doctoral dissertation, University of Pittsburgh]. ProQuest LLC.

Wiese, R. (1984). Language production in foreign and native languages: Same or different? In H. W. Dechert, D. Möhle & M. Raupach (Eds.), *Second language productions* (pp. 11–25). Tübingen: Narr Verlag.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*, 85-106.

Winke, P., & Lim, H. (2014). Effects of test wiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation, *IELTS Research Reports Online Series, 3*, 1-30. Retrieved from http://www.ielts.org/researchers

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

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency accuracy and complexity*. Honolulu, Hawaii: University of Hawaii at Manoa.

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*, 680-704.

Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interview. *Studies in Second Language Acquisition*, *14*, 403–424.

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

Zaki, H. and Ellis, R. (1999). Learning vocabulary through interacting with a written text. In R. Ellis, (Ed.), *Learning a second language through interaction,* (pp.153-169). Amesterdam: John Benjamins.

Zhou, Y. (2008). A comparison of speech samples of monologic tasks in speaking tests between computer-delivered and face-to-face modes. *JLTA Journal, 11*, 189-208.

Zhou, Y. (2015). Computer-delivered or face-to-face: Effects of delivery mode on the testing of second language speaking. *Language Testing in Asia, 5*, 1-16.

APPENDIX A

VOCI

1. My name is Gene and this is Ron. What is your name?

2. How long have you been in the United States?

3. Where do you live in the United States?

4. My favorite color is purple. Which of these colors do you like?

5. I am only wearing one color. What colors are you wearing today?

6. Ron is eating his dinner. What do you eat for dinner?

7. Gene loves desserts. What desserts do you like to eat?

8. What kinds of drinks do you like?

9.A: My wallet is almost empty, yours is so full

B: What do you have in your wallet?

10: What do you think she has in her bag? Name at least three things.

11. Name at least five things that are represented in this picture.

12. This is a picture of my hometown. Tell us about your hometown.

13. Instead of writing letters, you have decided to send a cassette message to a friend back home. Describe where you are living now and what you've been doing recently.

14. You have arrived on this campus and are approaching the information booth. Ask

15. You are planning your next vacation to Hawaii. You go to the travel agency to find out about the schedule and cost of flight, the cost of lodging, and the availability of tours. Ask the agent for this information.

16. Gene is packing her suitcase. What will she pack for her trip to Hawaii?

17. A. I'm so happy my best friend just got back from vacation. I really missed him a lot.

B: My best friend moved away and she's impossible to replace because she's so special.

A: Describe one of your friends.

18. Because of a last-minute problem you missed a dinner engagement with a friend. You call to apologize, but your friend is not yet home, so you need to leave a message on the answering machine apologizing for the date and explaining why you were not there.

19. A: Did you know that I went to New York last month? It sure is an interesting city.

B: What's so special about it?

A: The entire time I was there, I tried to compare it with our city. There are lots of differences, but on the other hand, lots of things are similar.

B: Can you compare your hometown with a city you visited or know very well?

20. A: One thing that I didn't like about New York was that it is so big. I never really feel comfortable in big cities anymore.

B: Why not? I love city life. There's nothing more fascinating than a really big city.

A: Not me. There are too many problems I guess.

B: What do you think? What are the advantages or disadvantages of big city life?

21. A: It's really unbelievable.

B: Yes, that was a really terrific experience.

A: There are some experiences you just can't forget.

B: That's true. Have you ever had such an experience—an experience that you'll never forget?

A: It can be something positive or it can be something negative.

B: Tell us about it.

22. A: So, you've finally made up your mind?

B: Yes, and I'm really excited about it.

A: Then you must have pretty concrete plans for the next few years?

B: Yes, and I also have a good idea about what my life might be like.

A: And you, what are your plans? What do you need to reach your goals?

B: How might your life look ten years from now?

23. You have a summer job selling great books, I am a potential consumer, convince me why should I buy the books from you.

24.A: Gene, did you read about the student who took a Swiss army knife to school in his pocket.

B: No, what happened?

A: Well, he was using a scissor part of it and his teacher caught him and she took it and they expelled him from school.

B: I don't get it; it looks like an innocent tool to me.

A: Well, their school has a zero-tolerance policy and they consider a Swiss army knife a weapon.

B: If you were the principal of the school, what would you do about this issue?

25. A: Wow look at the headline, another war.

B: There've always been wars, it is nothing new, it is just human nature.

A: Not necessarily.

B: How do you feel about this issue, how do you think we can create a lasting peace?

26. A: I really love this painting.

B: I don't understand it at all.

A: Tell us why you think this is or isn't art.

27. A: My computer is broken again.

B: Man, what a disaster!

A: Yeah, I feel so dependent on these machines.

B: Modern technology can make life easy, but it can cause a lot of frustration too.

A: Discuss the positive benefits and the negative consequences of our dependence on such machines.

28. A: Some undergraduates at American universities think that native speakers of English make the most effective teachers.

B: On the other hand, some people think the advantages of having an international teacher outweighs the disadvantages.

A: What do you think?

29. If you were a teacher and you discovered that one of your students had cheated on a test by copying from another student's paper, what would you do?

30.A: In many countries, higher education is for an elite group of students, not everybody can go to the university.

B: That's certainly not the case in this country. Our universities are open to everyone regardless of their background.

A: I can see the pros and cons for both types of educational systems.

B: Discuss the advantages and disadvantages of both educational systems.

31. A: I'm reading an article about free trade in Europe and in America, and it says that everybody benefits from having free trade.

B: Not really, there are still different positions in few countries about the whole issue of free trade.

A: Take one position and defend your opinion about free trade

32.A: Did you know that US law allows trials to be televised.

B: Yes, several high-profile trials have been televised recently because of the Freedom of Information Act.

A: I wonder if that's such a good idea.

B: What do you think about televising criminal trials?

33. A: Have you noticed how many shows on TV portray violent crime?

B: It's pretty hard not to notice!

A: Some people feel that this creates violence in our society.

B: Yes, but other people feel that it has no effect on young people. In fact, they're proud of this country's freedom of expression.

A: What do you think about the portrayal of violence and crime on TV?

34. A: there must be problems in your country, too.

B: What are some of the problems in your country?

A: Suggest some solutions and discuss the implications of these solutions.

35. This is the last question. If you've gotten this far, you probably have taken other English tests. If so, how does this test compare to the other English tests you have taken.

Appendix B

Online Background Survey

1. Choose a pseudonym for yourself and make sure you use it in all tests.
2. How old are you?
3. What class level are you in?
4. What is your academic major?
5. What is your native language?
6. How long have you been studying English, in years?
7. How long have you been in the United States?
8. What are some of the standardized tests you have taken before that tested your oral proficiency?
9. If you have taken more than one test, which test did get higher in the speaking section? (Q9)
10. How do the tests you have taken; the face to face interview and the video, compare to your testing experience with TOEFL and IELTS? Which one or ones do you prefer? (Q10)

VITA

Nawal Ali Alzahrani

Candidate for the Degree of

Doctor of Philosophy

Thesis: A COMPARATIVE STUDY OF ORAL PROFICIENCY IN DIRECT (OPI) AND SEMI-DIRECT (VOCI) TESTING MODES: MEASURES OF COMPLEXITY, ACCURACY, AND FLUENCY

Major Field: English/ Linguistics and TESOL

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Linguistics and TESL at Oklahoma State University, Stillwater, Oklahoma in May, 2020.

Completed the requirements for the Master Arts in Applied Linguistics at Umm Alqura University, Makkah, Saudi Arabia in 2010.

Completed the requirements for the Bachelor of Arts in English at Umm Alqura University, Makkah, Saudi Arabia in 2006.

Experience:

Graduate Teaching Assistant in the English department, Oklahoma State University, United States, 2018-2020

Writing Consultant in the Writing Center, Oklahoma State University, United States, 2018-2020

Part-time language instructor in the English Language Institute at Oklahoma State University, United States, 2014-2019

Lecturer at the English Language Center, Umm Alqura University, Saudi Arabia, 2010-2013

Arabic-English Interpreter, National Guard Hospital, Saudi Arabia, 2009-2010

Part-time language instructor in the English department, Umm Alqura University, 2007-2009

Professional Memberships: American Association of Applied Linguistics (AAAL) / Arabic Linguistic Society (ALS)
Linguistic Society of America (LSA)/ Teachers of English to Students of Other Languages (TESOL)