PREDICTING INTERNATIONAL TEACHING ASSISTANTS' PERFORMANCE IN A DOMAIN-SPECIFIC TEST: THE CASE OF COMPLEXITY, ACCURACY, FLUENCY, AND COMPENSATORY STRATEGIES

By

SHAHRIAR MIRSHAHIDI

Bachelor of Arts in English Translation Training
Payam Noor University
Mashhad, Iran
2007

Master of Arts in Teaching English as a Foreign Language
Islamic Azad University, North Tehran Campus
Tehran, Iran
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, 2017

PREDICTING INTERNATIONAL TEACHING ASSISTANTS' PERFORMANCE IN A DOMAIN-SPECIFIC TEST: THE CASE OF COMPLEXITY, ACCURACY, FLUENCY, AND COMPENSATORY STRATEGIES

Dissertation Approved:

Dr. Gene B. Halleck

Dissertation Adviser

Dr. An Cheng

Dr. Stephanie Link

Dr. Kenneth Clinkenbeard

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my adviser, Dr. Gene B. Halleck, whose mentorship not only made me a better applied linguist, but it also made me a better human being. She taught me how to judge my test takers' language ability but not to judge them by their skin color, their first language, or their religion. She helped me get involved with the International Teaching Assistant Program at Oklahoma State University from the very beginning of my doctoral education. Without her invaluable guidance and help completing this dissertation work would not have been possible.

I am also very grateful to my committee members, Dr. An Cheng, Dr. Stephanie Link, and Dr. Ken Clinkenbeard. Being mentored by Dr. Cheng taught me that one does not need to be a native speaker to have an influential academic presence in the field of Applied Linguistics. Also, Dr. Link's suggestions and input helped me a lot in the process of finishing my dissertation project. I hereby thank the efforts of my other professors, Dr. Dennis Preston and the late Dr. Ravi Sheorey, in shaping my identity as a researcher.

I am endlessly indebted to my dear friend and old companion, Dr. Manoucher Motamedi whose inspirational support has always steered me in the right direction in life. Furthermore, I express my gratitude to my dear friend and colleague, Dr. Hooman Saeli, for his constructive and precious peer feedback along the way.

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

I am thankful to my father, my mother, and my sister for their unconditional love and support. I specifically thank my mother who sent me to my first language school as a learner at the age of seven when I had no idea that English would one day become my passion, my weapon, and my future path. I also thank my Bita for being there for me even when I was tired, agitated, hopeless, and cranky. For certain, her love made me more determined to go on.

Finally, I have to extend my admiration and respect to all the ITAs whom I had the pleasure to serve either as a tester or as an educator. Particularly, I take my hat off to the Iranian ITAs who currently teach under tough conditions in the United States where the rhetoric of hatred and xenophobia is constantly echoed.

Shahriar Mirshahidi
March 21, 2017

Name: SHAHRIAR MIRSHAHIDI

Date of Degree: MAY 2017

Title of Study: PREDICTING INTERNATIONAL TEACHING ASSISTANTS' PERFORMANCE IN A DOMAIN-SPECIFIC TEST: THE CASE OF COMPLEXITY, ACCURACY, FLUENCY, AND COMPENSATORY STRATEGIES

Major Field: ENGLISH (TEACHING ENGLISH AS A SECOND LANGUAGE)

Abstract: International teaching assistant (ITA) assessment for screening purposes has gained crucial importance with the increase in the number of nonnative speaker graduate students across North American campuses. Although previous studies show that ITA proficiency entails different skills from general oral proficiency, research on its operationalization is rather scarce. Furthermore, while complexity, accuracy, and fluency (CAF) have proven to be reliable constructs to describe language performance (e.g. Ellis & Barkhuizen, 2005; Housen, Kuiken, & Vedder, 2012; Pallotti, 2009), they have not found their way to research on ITA proficiency. Additionally, communication strategies can help ITAs compensate for their linguistic shortcomings. In particular, Halleck and Moder (1995) suggest that employment of compensatory strategies may help ITAs meet the standards that they require in order for undergraduate students to comprehend their oral speech. To that end, more than three hours of ITA Test performances of the candidates at a southcentral university in the United States were videotaped. The ITA Test was an in-house, domain-specific performance test for screening ITA candidates at the university. Using a validated, holistic rubric, the trained raters assigned the performances to three categories of Passed, Provisionally Passed, or Failed. Then, 21 performances were selected and analyzed incorporating eight measures of CAF and two measures of compensatory strategies. The findings revealed that accuracy, when measured by percentage of error-free T-units and number of unintelligible words per 100 words, predicted ITA proficiency. With regard to compensatory strategies, using nonlinguistic means (i.e. visual aids, eye contact, and body language) also predicted the ITA candidates' performance. To sum up, error-free T-units, nonlinguistic means, and number of unintelligible words per 100 words predicted ITA candidates' oral performance when assessed by the ITA Test. Finally, the trade-off effects between the CAF measures (i.e. between accuracy and complexity and between accuracy and fluency) partially confirmed Skehan's (1998) Limited Capacity Hypothesis. Overall, the results can shed light on the role linguistic and paralinguistic features play in predicting ITAs' test performance for both assessment and pedagogical purposes.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION 1

1.1. International teaching assistants (ITAs) in the States 1
1.1.1. The foreign TA problem 1
1.2. ITA Proficiency: A new perspective...2
1.2.1. L2 speaking proficiency 2
1.2.2. ITA proficiency: An independent construct 3
1.3. Complexity, accuracy, and fluency (CAF) and language Performance............5
1.3.1. Complexity...6
1.3.1.1. Measures of complexity..8
1.3.2. Accuracy 10
1.3.2.1. Measures of accuracy...10
1.3.3. Fluency...11
1.3.3.1. Measures of fluency...12
1.3.4. The relationship between CAF constructs 13
1.3.5. CAF and ITA proficiency: The missing link 14
1.4. ITAs and the use of compensatory strategies 14
1.5. This study and statement of the problem 16

II. LITERATURE REVIEW...19

2.1. Overview..19
2.2. Research on ITA education and assessment 19
2.3. Research on CAF constructs...24
2.3.2. Research on complexity...28
2.3.3. Research on accuracy...32
2.3.4. Research on fluency..36
2.4. Research on compensatory strategies and communicative adequacy.............42
2.5. The missing link..47

III. METHODOLOGY 49

3.1. Overview...49
3.2. Participants..50
3.3. Context of the study...51

3.4. Instruments and variables 52
3.4.1. The ITA Test..........52
3.4.2. The ITA Test raters..........54
3.4.3. The rating process and the rubric..........55
3.4.4. Operationalization of the variables..........57

Chapter Page

3.4.4.1. Measuring complexity 57
3.4.4.2. Measuring accuracy..........60
3.4.4.3. Measuring fluency 63
3.4.4.4. Compensatory strategies..........65
3.5. Procedures and data collection..........67
3.6. Statistical analysis..........69

IV. RESULTS..........71

4.1. Overview..........71
4.2. Normality of the data 72
4.3. Descriptive statistics 73
4.4. RQ1: Predictive power of CAF constructs 73
4.4.1. Result of the multiple regression analysis 75
4.4.2. Result of the post-hoc analysis 78
4.5. RQ2: The relationship between CAF constructs 79
4.6. RQ3: Predictive power of compensatory strategies..........80
4.6.1. Result of the multiple regression analysis 81
4.6.2. Result of the post-hoc analysis 82
4.7. RQ4: Comparing CAF constructs and compensatory strategies..........83
4.7.1. Comparing the obtained R^2 statistics..........83
4.7.2. Comparing the obtained β values..........84
4.8. Summary of the results 84

V. DISCUSSION..........86

5.1. Overview..........86
5.2. Discussion of the answers to the research questions 87
5.2.1. Predictive power of CAF constructs..........87
5.2.2. Competition between CAF constructs 89
5.2.3. Predictive power of compensatory strategies 91
5.2.4. Comparing CAF constructs and compensatory strategies 94
5.3. General discussion and testing/teaching implications 95
5.4. Limitations of this study and future directions..........98
5.5. Conclusion 100

REFERENCES 102

APPENDICES 117

LIST OF TABLES

Table Page

1 Error categorization of Weighted Clause Ratio 35
2 Typology of communication strategies (Yaman & Özcan, 2015) 43
3 Decisions on international graduate students’ applications for TAship 51
4 The ITA Test characteristics based on Bachman's test method facets 53
5 Interpretation of the ITA Test scores..56
6 Summary of CAF measures used in this study...65
7 Descriptors of the score on nonlinguistic means 67
8 Inter-coder reliability of the coded measures 68
9 Results of the test of normality..72
10 Descriptive statistics of the measures across the three ITA groups....................74
11 Significant correlations between CAF measures 80

LIST OF FIGURES

Figure Page

1 Bulté and Housen’s (2012) taxonomy of L2 complexity......................................30
2 Vercellotti and Packer’s (2016) proposed order of clause development 32
3 Typology of speech fluency (Based on de Jong et al., 2015; Segalowitz, 2010) ..37
4 Skehan's (2014) framework of speech fluency 38
5 The ITA Test structure and sequencing 54
6 The means for PEFT (a) and NUIW (b) across the three ITA test groups 76
7 The scatterplot of ITA Test scores for PEFT 76
8 The scatterplot of ITA Test scores for NUIW 77
9 The means plot for PEFT 78
10 The means plot for NUIW 79
11 The means for NLM across the three ITA test groups...81
12 The means plot for NLM 82

CHAPTER I

INTRODUCTION

1.1. International teaching assistants (ITAs) in the States

International students applying to graduate schools in the United States play a major role in shaping the structure of many graduate programs across the country. A report published by Council of Graduate Schools (CGS, 2013, November 5) indicates that between 2012 and 2013, the number of enrollments amplified by 10%. This increase was recorded in almost all fields of study with Physical and Earth Sciences (18%) and Engineering (17%) experiencing the highest boom. Despite a more recent report by CGS that shows this rise in the number of international applicants has slowed down, these foreign-born students are still substantial to the existence of their respective graduate programs, particularly in technical fields (Chiang, 2009). Moreover, to cover for the expenses of graduate school, some of these students serve as teaching assistants (TAs). Graduate programs also welcome these international students occupying TA positions, since they fill in for the Americans who leave college in order to pursue occupational goals (Inglis, 1993).

1.1.1. The ‘foreign’ TA problem

Most of the academic institutions in the United States (both public and private) now deal with screening, admitting, and training ITAs. More importantly, this involvement not only includes program directors and school officials, but it also comprises parents and even state legislatures (Bailey, 1983, 1984; Bresnahan & Cai, 2000; Halleck & Moder, 1995; Yule & Hoffman, 1990). Due to the involvement of too many stakeholders, concerns have grown over the language proficiency of these ITAs (Halleck, 2008; Kang, Rubin, & Lindemann, 2015), and as a result, both assessing the proficiency of and training ITA candidates have gained the utmost significance (Choi, 2017). Numerous studies and official reports point out that linguistically incompetent ITAs would jeopardize the undergraduate students' chances of succeeding in the courses that involve ITAs to a great extent. This adverse impact may continue to the level that it affects the decision-making process of undergraduates whether or not to pursue education in their fields (see Seymour & Hewitt, 2000). It is not difficult to decipher that many programs that use ITAs will not be able to exist without them. However, the pressure is mounting on such programs to take sticter measures when screening and employing ITAs (e.g. International Herald Tribune, 2005, June 25). Here, questions are raised about that how much a graduate program can rely on ITAs to handle instructional duties and in what way should we expect ITAs to fulfill these duties. Moreover, what are the constituents of these duties? How much of what we should look for in an ITA is related to language? Finally, what components of language contribute to ITAs' success in meeting the expectations?

1.2. ITA Proficiency: A new perspective

1.2.1. Second language (L2) speaking proficiency

Within an L2 domain, and in a very general sense, proficiency may be defined as the ability of a second language user to perform and function in that language in real-life settings (Thomas, 1994). According to Iwashita (2010, p. 32), "four key traits" of lexical diversity, grammatical accuracy, syntactic complexity, and fluency establishes second language proficiency. In the same vein, L2 speaking proficiency deals with the ability of the learner to interact in authentic situations using the spoken modality of L2. Nonetheless, a lack of consistency in the chosen measures and the ways different researchers operationalize the construct of speaking proficiency make the generalizability of L2 oral proficiency assessments rather questionable (Bowden, 2016). The range of subconstructs that significantly influence speaking proficiency are reported to include vocabulary and grammar at lower levels and fluency and sociocultural factors at higher levels (e.g. Adams, 1980), pronunciation at lower and fluency at higher levels of proficiency (e.g. de Jong & van Ginkel, 1992), length and complexity (e.g. Tamaru, Yoshioka, & Kimura, 1993), task type and proficiency level (e.g. Halleck, 1995, Iwashita, 2006), task design (e.g. Tavakoli & Foster, 2011), and vocabulary and fluency (e.g. Iwashita, Ortega, Rabie, & Norris, 2008).

Furthermore, lack of consistency in the operationalization of proficiency has led many researchers to merely rely on self-reported evaluations, given proficiency levels at the institutional level, or subjective judgments instead of “either standardized or in-house assessments” (Bowden, 2016, p. 4). But even when the oral assessment section of standardized tests is used as the basis to research the contributing factors to L2 oral proficiency, the results may not be generalizable enough to predict the true and specific abilities that an L2 user might need in particular contexts. That is, as Norris and Ortega (2012, p. 580) argue, a “major limitation to the generalizability” of research findings in this area can be scarcity or absence of “domain-specific” language proficiency tests.

1.2.2. The independent construct of ITA proficiency

Many traditional tests of English oral proficiency have been used across different U.S. campuses to screen L2 communicative abilities of ITA candidates, such as the Oral Proficiency

Interview (OPI), Spoken Proficiency English Assessment Kit (SPEAK), and Test of Spoken English (TSE). More recently, the speaking modules of popular proficiency tests like the internet version of Test of English as a Foreign Language (TOEFL) and International English language Testing System (IELTS) are used to evaluate, and therefore predict, the ITA candidates' ability to teach the subject matter in English in an American context. However, there is no agreement among researchers in the field on how much current, widely-accepted theories of L2 assessment (in particular, those aspects related to oral proficiency) which are implemented in construct validation of standardized tests are in line with the nature of ITA proficiency. Ard (1987), for instance, believes that such theories fail to account for the ITA situation in that they establish the kind of language learning which relies more on formal aspects of language (e.g. grammar and pronunciation), whereas, in reality, teacher discourse and strategy usage are also integral to ITA education. In her study, Elder (1993) found a poor correlation between IELTS scores (both overall scores and speaking scores) and teaching performance of international students in a teacher training course, and therefore, questioned the construct validity of IELTS to assess this domain-specific proficiency. Thus, researchers of the field have tried to define and operationalize ITA proficiency as a *separate construct* that has both similar and distinct features to and from L2 oral proficiency (McGregor, 2007; Mirshahidi & Saeli, 2016; Saeli & Mirshahidi, 2014).

Arguably, what constitutes L2 oral proficiency does not entirely suit the qualities that an ITA need to exhibit in the classroom or lab. In other words, ITAs perform in a "specific context" which requires "necessary skills" for their competence to be demonstrated (Halleck & Moder, 1995, p. 734). The importance of authentic (or *authenticated*) context is often emphasized in the literature on ITA proficiency assessment. Research indicates that assessing an L2 for specific purposes is a better indicator of success in related, future performances compared to tests of overall language proficiency (see Douglas & Selinker, 1992; Fulcher, 2015; Norris and Ortega, 2012). Therefore, it can be concluded that ITAs' performance is composed of "a complex activity which requires the employment of all aspects of communicative competence" (Saif, 2002, p. 147). Nevertheless, as McGregor (2007) claims, there is no consensus on the type and scope of *additional* components relevant to ITA assessment and training, including teaching skills and sociocultural competencies.

1.3. Complexity, accuracy, and fluency (CAF) and language performance

CAF measures (also, CAF dimensions or CAF constructs) have long been used in second language acquisition (SLA) research to track the development of second language learners. Mostly, CAF measures have been used to describe learners' performance in productive skills (i.e. oral and written production). To be specific, these measures are "dimensions for describing language performance" with regard to variable features such as task characteristics and learner differences (Pallotti, 2009, p. 590). Also, Skehan (2009) describes CAF as "useful measures of second language performance" (p. 510). Furthermore, Housen and Kuiken (2009, p. 461) describe CAF measures as "indicators of learners' proficiency underlying their performance." They argue that complexity and accuracy *represent* learners' knowledge of the L2, whereas fluency *reflects* their mastery over this knowledge (p. 462). The body of research on CAF constructs and their various aspects is abundant. In the next section, I will define each construct based on the available literature. It is worth mentioning that defining CAF, measuring its three constructs, and the relationship between them have not been without controversy (Housen & Kuiken, 2009; Larsen-Freeman, 2009; Vercellotti, 2012). This issue will be discussed in detail in the upcoming sections.

1.3.1. Complexity

Pallotti (2009, 2015) believes that complexity is the most difficult construct to define among CAF measures due to its multidimensional nature. Housen and Kuiken (2009) describe complexity as “the most complex, ambiguous, and least understood dimension of the CAF triad” (p. 463). Complexity should be approached with caution, since the term is used in the literature to refer to both an independent variable of task complexity (see Tavakoli & Foster, 2011) and a dependent variable of linguistic or structural complexity (see Pallotti, 2015). Even linguistic complexity is multifaceted; that is, it comprises a form of complexity that deals with L2 learners’ interlanguage system, and therefore, has a developmental nurture (Housen & Kuiken, 2009; Pallotti, 2009). In other words, as the learner matures at the interlanguage stage, the language that is produced becomes richer and more elaborate. More commonly though, linguistic complexity is conceptualized as a constituent of language performance as it is produced by the L2 user (Pallotti, 2009). Pallotti (2015, p. 118) categorizes three main meanings attributed to complexity in the literature:

1. “Structural complexity, a formal property of texts and linguistic systems having to do with the number of their elements and their relational patterns;
2. Cognitive complexity [or *difficulty*], having to do with the processing costs associated with linguistic structures;
3. Developmental complexity, i.e., the order in which linguistic structures emerge and are mastered in second (and, possibly, first) language acquisition.”

As can be seen above, there is no consistency in defining this very construct (Bulté & Housen, 2012, 2015; Housen, Kuiken, & Vedder, 2012). As Pallotti (2009) and Housen et al. (2012) argue, both cognitive and developmental notions of complexity are dynamic, because they often differ across learners (i.e. based on variables such as, motivation and age) and are subjectively measured. Nonetheless, structural complexity is "a more stable property of the individual items, structures or rules" than can be observed, recorded, and objectively measured in L2 users' language production (Bulté & Housen, 2012, p. 25). Some researchers (e.g. Ellis, 2009; Ellis & Barkhuizen, 2005; Skehan, 2009) view complexity as the ability of the learner to produce more elaborate and advanced language. However, there is no agreement in the findings of related studies on what precisely constitutes advanced and elaborate language. On the contrary, a more in-depth, multilayered perspective of structural complexity has achieved more agreement on the side of some CAF researchers. This group of researchers suggest that complex language can be viewed to have multiple elements, layers, and forms within its structure (see Bulté & Housen, 2015; Pallotti, 2015).

It seems that a thorough way to investigate complexity would be to adopt a hybrid definition based on the research objectives and research questions which are also empirically tested and grounded in theory. Thus, in this study, I define complexity as a construct that, along with accuracy and fluency, is used to describe, predict, or explain L2 performance as well as L2 proficiency. In this view, which is close to that of Pallotti's (2015), complexity is defined as the number of components (or layers) that make up a selected grammatical unit and the relationship between these components. I have to emphasize that this conceptualization does not drastically deviate from the existing definitions of complexity that are already in the literature on CAF. Besides, Pallotti (2015, p. 120), claims that, within structural complexity, a difference should be made between *grammatical* complexity (i.e. “complexity of grammatical rules”) and *stylistic* complexity (i.e. complexity influenced by learners’ “culture-specific rhetorical patterns”). However, for research purposes, it is rather unfeasible to identify all these culture-specific patterns that are writer’s or speaker’s stylistic choices, specifically when the researcher deals with L2 learners from multiple language backgrounds. In addition, there is no research-informed fine line between cultural-driven patterns in using complex language and complexity in the grammatical sense (see Johnson, 2004).

Structural complexity can also be narrowed down into three subcategories including syntactic complexity, lexical complexity, and morphological complexity. In the present study, I only used syntactic complexity (i.e. number of grammatical components in a chosen unit of produced language and the relationship between these components) along with the other CAF constructs. The rationale behind this decision is discussed in detail in the Methodology.

1.3.1.1. Measures of syntactic complexity

Measures that are used in research on CAF to assess syntactic complexity are mostly divided into two groups of *general* and *specific* measures. General measures are usually confined to measures of length and subordination (for an extensive review of all complexity measures, see Bulté & Housen, 2012).

Length-based measures use the ratio of frequency of words to the total number of the chosen syntactic unit. A very popular measurement unit in SLA research is minimal terminable unit or T-unit (Norris & Ortega, 2009). A T-unit may be defined as “an independent clause and all of its dependent clauses” (Iwashita, 2006, p. 157). Previous research on CAF and oral proficiency has testified on the functionality of T-units in predicting proficiency levels (e.g.

Halleck, 1995; Iwashita, 2006). There are also other units of measurement such as Analysis of Speech Unit (AS-unit) which has been utilized in some recent studies concerning oral proficiency (e.g. Révész, Ekiert, & Torgersen, 2014; Tavakoli, Campbell, & McCormack, 2016). Despite popularity of T-unit, Foster, Tonkyn, and Wigglesworth (2000) warn against the usefulness of such a unit of measurement for spoken data. They claim that dysfluencies and broken language cannot be properly measured through T-units. However, the definition they provide for their suggested unit (i.e. *AS-unit* as the "single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either", Foster et al., 2000, p. 356) does not radically differ from the cited definitions of T-unit in the literature. More importantly, such disfluencies that tend to exist in conversations and impromptu talk may not appear in ITA discourse, since, for the most part, it involves preplanning and is mostly monologic (The logic behind choosing T-unit as the unit of analysis for spoken language in this study is discussed more in Chapter III).

Subordination is another general measure to investigate the construct of complexity. It is normally "computed by counting all clauses and dividing them over" the chosen production unit (Norris & Ortega, 2009, p. 558). For instance, dividing the number of both dependent and independent clauses over the number of T-units within a pre-determined chunk of spoken data can be used as a general index of syntactic complexity. Alternatively, though, the mean number of T-units or AS-units have also been used (Iwashita, 2006).

Specific measures are utilized for more narrowed-down research objectives. These measures essentially include tokens of certain syntactic forms (i.e. the number of tokens per, for example, 100 words). Such tokens may include tense-aspect forms, modal verbs, or types of clauses, for example, additive, temporal, relative, and so forth (Révész et al., 2014).

1.3.2. Accuracy

Unlike complexity, accuracy is the most agreed upon construct of CAF (Housen et al., 2012; Pallotti, 2009). Accuracy essentially refers to the production of error-free language. In other words, speech that complies with the linguistic "norms" of the "target" language will be regarded as accurate (Yuan & Ellis, 2003, p. 2). Since the terms *error* and *norm* themselves are problematic in nature, researchers need to be cautious about how they treat errors. According to Housen et al. (2012, p. 4), who define errors as "deviations" from norms, such nonconformities should be contextually identified before accuracy is assessed. Deviations from the accepted linguistic norms across certain contexts (e.g. an academic monologic presentation vis-à-vis a casual conversation) are different considering all the intervening factors, such as communicative goals and audience. Therefore, *appropriateness* of the produced language is complementary to its accuracy.

It is of particular note that accuracy may cover both notions of *grammatical accuracy* and *lexical accuracy* (Lennon, 1995; Tonkyn, 2012). While grammatical accuracy includes appropriateness and conformity with norms of forms and syntactic rules (e.g. *A part of civil engineering deals <u>about</u> structural analysis.*), lexical accuracy refers to the extent to which the lexicon in the produced language is chosen appropriately and within the accepted norms of the respective context (e.g. *I'm going to talk about what consumer surplus means and how to <u>calculate this concept</u>*).

1.3.2.1. Measures of accuracy

Similar to complexity, there are general and specific measures available to researchers to quantify and evaluate accuracy in L2 production. One of these measures is achieved through dividing the total number of errors by the total number of words, especially for a short selection. Instead, when analyzing longer stretches of language, the number of errors per, for example, 100 words, can be calculated. Depending on research scope and objectives, errors may include both errors in grammar and lexis (see Révész et al., 2014). Additionally, the number or proportion of error-free units (e.g. error-free T-units) is another prevalent general measure of accuracy (Tonkyn, 2012). Foster and Skehan (1999) argue that percentage of error-free clauses is a reliable and sensitive measure of accuracy (see also Iwashita et al., 2008). In addition, Tonkyn (2012) warns against operationalizing comprehensibility and accuracy as independent constructs arguing that "attempts to an error gravity hierarchy have produced conflicting results" (p. 225).

Finally, accuracy in L2 performance can also be assessed by looking at specific errors, such as tense-aspect errors, errors in subject-verb agreement, erroneous active or passive voice, incorrect modal verbs, and so forth (see Foster & Wigglesworth, 2016). One way to approach accuracy through specific measures would be to investigate over-suppliance and under-suppliance of the above-mentioned examples in the learner's production.

1.3.3. Fluency

Pallotti (2009) argues that, like complexity, fluency is a multi-faceted construct. Tavakoli et al. (2016) suggest that speech fluency not only incorporates linguistic features but it also integrates social and psychological aspects of language. Lennon (1990, 2000) states that in a 'broad' sense, which is nontechnical, fluency is used interchangeably to refer to oral proficiency. However, in his description of 'narrow' sense, fluency is distinguished from other proficiency constructs (e.g. accuracy and complexity) and is characterized by speech rate, pausing, hesitation, and repairs. By this definition, fluency is "purely a performance phenomenon" that heavily relies on the listener's role in evaluating this performance (Lennon, 1990, p. 391).

Skehan (2009, p. 510) defines fluency as “the capacity to produce speech at normal rate and without interruption.” In Richards and Schmidt’s (2010) viewpoint, fluent speech is natural in a sense that pauses, speech rate, and intonational features are close to the target language norms. One factor to consider here is the potential trade-off effect between fluency and the other measures, that is, complex and accurate performance might come at the price of a less fluent and more hesitant speech (Skehan, 1998, 2009; Vercellotti, 2012, 2015).

There are various classifications of fluency and its different aspects. Among these classifications, categorizing fluency into *cognitive fluency*, *utterance fluency*, and *perceived fluency* has gained research significance (see Segalowitz, 2010). While cognitive fluency deals with the unobservable cognitive processes involved in producing speech and perceived fluency is about listener’s evaluations toward the speech they hear, utterance fluency denotes “the measurable aspects of fluency such as speed, pausing, and hesitation” (Tavakoli et al., 2016, p. 449). In this study, fluency exclusively refers to the concept of utterance fluency.

1.3.3.1. Measures of fluency

Utterance fluency itself can be regarded as having three different subconstructs: speed, breakdown, and repair (Révész et al., 2014; Skehan, 2003, 2014; Tavakoli et al., 2016). Each of these fluencies can be quantified and assessed by using several measures. Speed fluency that measures the velocity of speech is quantifiable through articulation rate and mean syllable duration (de Jong, Steinel, Florijn, Schoonen, & Hulstjin, 2013). Articulation rate can be computed by dividing the number of syllables by total speaking (or phonation) time, and mean syllable duration is calculated by the inverse computation of articulation rate (de Jong, Groenhout, Schoonen, & Hulstijn, 2015). Breakdown fluency embodies silence and pausing in speech (Tavakoli et al., 2016). Numerous measures have been proposed to weigh silence and pausing in speech. Number of filled pauses or number of silent pauses per 100 words are among popular measures of breakdown fluency (Révész et al., 2014; Skehan, 2009). Also, computing mean length of silent pauses within and between T-units or AS-units is another way to measure breakdowns in speech fluency (de Jong et al., 2015). Following de Jong et al. (2013), Révész et al. (2014) suggest that the cut-off point for silent pauses should be set at 250 milliseconds. Repairs in fluency usually includes false starts, self-repairs (or reformulations), and repetitions. In their study on communicative adequacy, Révész and her colleagues used the ratio of number of tokens (from each type of repair fluency) by 100 words. In addition, repair measures may include the mean number of complete/incomplete repetitions, false starts, or reformulations in a particular amount of speech, for instance, per 60 seconds (Tavakoli et al., 2016).

Selection of fluency measures needs to be done with care, since there is a possibility of “overlap” between what they are supposed to quantify (Tavakoli et al., 2016, p. 456). For instance, some speech fluency measures, such as speech rate (a speed fluency measure calculated by dividing the number of syllables by total speech time including pausing time), may provide the researchers with mixed results which do not distinguish the exact magnitude of the impact caused by pausing and/or speed (see de Jong et al., 2013; Skehan, 2014).

1.3.4. The relationship between CAF constructs

Many studies suggest that there are limitations to the level of cognitive load that is required for performance in L2. In other words, there is a competition between CAF constructs in a way that focusing on one of the components (e.g. on accuracy) may result in poorer performance in the other (e.g. in fluency). Skehan (1998) explains this competition in the form of his Limited Capacity Hypothesis that proposes a *trade-off effect* between linguistic components of performance due to limited attentional capacity and working memory. Perhaps a rival view to that of Skehan is Robinson’s (2003, 2005) Cognition Hypothesis that emphasizes the role that task characteristics play in shaping the cognitive demands that ultimately spark rivalry between CAF components. Simply put, learner’s performance becomes more accurate and complex as the task becomes more complex, that is, more difficult (Robinson, Cadierno, & Shirai, 2009). Dynamic systems theory or DST (de Bot, 2008) and complexity theory (Larsen-Freeman & Cameron, 2008; Larsen-Freeman, 2009) inspire a third and most recent view on the relationship between CAF constructs. Considering these two similar theories, Vercellotti (2015) argues that, within CAF, “specific trade-off effects may be found, but they are not understood to have a causal, linear, or mutually exclusive relationship” (pp. 2-3).

1.3.5. CAF and ITA proficiency: The missing link

Although the body of research on measures of complexity, accuracy, and fluency (e.g. Ellis & Barkhuizen, 2005; Housen & Kuiken, 2009; Pallotti, 2009) verifies their validity and reliability in assessing oral language performance, these variables seem to have not found their way to ITA assessment. Without a doubt, analyzing the linguistic aspects of ITA proficiency based on CAF constructs can be a valuable addition to the existing studies that have tried to explain domain-specific, L2 performance relying on a componential view of language.

1.4. ITAs and the use of compensatory strategies

Studies on ITA training and assessment in the past several years have argued over a number of influential factors in ITAs’ success in both screening tests and actual classroom or laboratory teaching assignments (For a review of these studies see Farnsworth, 2013). Indisputably, for an ITA, an accurate and complex or a fluent and accurate performance does not necessarily guarantee that communication takes place effectively and the instructional objectives are met. Theories of communicative competence, in a general sense, assert the significance of strategic competence for effective communication (Bachman & Palmer, 1996; Canale, 1983). A number of studies have also stressed the importance of compensatory strategies as a major constituent of strategic competence (e.g. Celce-Murcia, Dörnyei, & Thurrell, 1995; Dörnyei & Scott, 1997; Natakatani, 2006). Such strategies can prolifically assist ITAs to compensate for their linguistic shortcomings and pass as successful teaching assistants (see Bailey, 1984; Halleck & Moder, 1995). Relatedly, Halleck and Moder (1995) suggest that employment of compensatory strategies may help ITAs meet the standards that they required in order for undergraduate students to understand them. They claim that "linguistic difficulties" will be less perceptible with a "clear and well-organized" presentation of teaching materials (p. 753).

Communication strategies that help ITAs compensate for their linguistic problems may include using visual aids (e.g. Microsoft PowerPoint™ or Prezi™ presentations, handouts, etc.), paralinguistic strategies (e.g. gestures and eye contact), and elaboration. ITAs can also employ certain strategies in question and answer (Q&A) situations, such as asking for clarification or postponing the answer to an unfamiliar question to a later class or time. Zha (2006) uses Cohen's (1998) classification of strategies to explain the strategies that can be taught to ITAs. Zha argues that cover strategies can be used by ITAs to handle situations in which the required linguistic knowledge is deficient to answer difficult queries or questions irrelevant to the content being instructed. Furthermore, using circumlocution or gap fillers can also be considered as learning strategies (Oxford, 2003) that can also be used as ITA compensatory strategies.

One final note would be that adopting compensatory strategies in ITA performance is not without its fallacy. Although many studies argue for the primacy of compensatory strategies, using such strategies, as Halleck and Moder (1995) claim, can only help those ITA candidates who are above a certain level of proficiency below which employment of strategies does not necessarily yield satisfactory results. That is to say, training courses that introduce and instruct compensatory strategies seem to benefit only ITA candidates who have higher proficiency levels.

1.5. This study and statement of the problem

As I discussed at the outset of the chapter, the increase in the number of international applicants to graduate schools in the U.S., and consequently occupying more TA positions by these students, ITA screening and training has gained a crucial momentum. Previous research on L2 oral proficiency, however, falls short in operationalizing the construct of ITA proficiency in order to assess thoroughly what constitutes the specific communicative abilities required in this domain. Thus, there has always been a need to develop and administer a test in which the target domain of abilities being tested are defined in such a way that making "accurate inferences about a test-taker's [in this case, ITA candidates] future performance" is feasible to a reliable extent (Elder. 2001, p.150).

Drawing on the results of the studies on CAF constructs and strategic competence, it is reasonable to hypothesize that these measures might also be a constituent of the construct of ITA proficiency. Surprisingly, to date, little research has been conducted utilizing CAF measures to analyze, explain, or assess ITA proficiency, although there is an abundance of literature on these measures in SLA research and oral language teaching and testing. Moreover, at the nonlinguistic level, a few studies on ITA education acknowledge that compensatory strategies can help ITAs compensate for their linguistic shortcomings (e.g. Ard, 1987; Halleck & Moder, 1995). Accordingly, many ITA programs have included instruction of these strategies in their curricula. It has been reported that in some of these programs, due to limitations in time and budget, most activities are devoted to teaching such strategies rather than linguistic issues and teacher discourse (Elder, 2001, Halleck & Moder, 1995). Nevertheless, there is lack of empirical research on the extent of the role that either linguistic components or compensatory strategies play in ITAs' success.

I argued earlier that ITA discourse, for the most part, involves preplanning and is predominantly monologic. Hence, in this study, I compared the most frequently used measures of CAF constructs and compensatory strategies in the monologic performance of the ITA candidates. Specifically, I focused on investigating the extent to which success (i.e. the dependent variable) in the monologic task of a valid ITA performance test could be predicted by measures of CAF and compensatory strategy use (i.e. the independent variables). At this point, I posed the following research questions that were partially inspired by Iwashita (2006), Révész et al. (2014), and Vercellotti (2012, 2015):

RQ1: Do CAF constructs and their related measures predict ITA candidates' performance in an ITA test? If so, what is the extent to which CAF constructs differentiate between ITA candidates based on their test performance?

RQ2: Is there a relationship between CAF constructs? If yes, how do CAF constructs influence each other when the task condition is not a variable?

RQ3: Do compensatory strategies predict ITA candidates' performance in an ITA test? If so, to what extent do compensatory strategies differentiate between ITA candidates based on their test performance?

RQ4: Upon comparing CAF measures and compensatory strategies, which one is a stronger predictor of ITA candidates' performance in the ITA test?

Throughout the next chapter, in a respective fashion, I will outline a review of the studies on ITA assessment and education, CAF constructs, their measures, and the relationships between these measures, along with a synthesis of research on compensatory strategies and communicative adequacy. These studies have formed theoretical and empirical bases to address the research questions posed above.

CHAPTER II

LITERATURE REVIEW

2.1. Overview

In Chapter I, I explained the critical situation of international teaching assistants (ITAs) in the United States. Moreover, I emphasized the role that complexity, accuracy, and fluency (CAF) play in describing learners' L2 oral performance. In addition, I briefly explained the nature of compensatory strategies as assistive tools available to L2 learners to compensate for their linguistic shortcomings. Finally, with regard to CAF and compensatory strategies, the objectives of this dissertation project, as well the related research questions were introduced. In the current chapter, I will go over the theoretical and empirical research, which has been conducted surrounding the variables of this study. A synthesis of both the old and the most recent studies on these topics will provide the readers with the opportunity to recognize the essential role that each variable can potentially play in validating the construct of ITA proficiency.

2.2. Research on ITA education and assessment

With the increase in the number of foreign-born graduate students who are employed as teaching assistants in American universities, concern has grown over the English language proficiency they possess. This concern has been around for quite a few decades now and it involves school officials, program directors, and even local undergraduate students who sit in classes or lab sessions run by ITAs. To date, many studies have tried to focus on this issue by investigating various aspects of ITAs' linguistic situation both before and during employment.

In one of the earliest articles published on ITAs in the United States, Ard (1987) warned that there are differences between normal second language acquisition and the type of language learning that ITAs need. He argued that ITAs' competence goes beyond language features such as morphology and syntax and labels it as "discursive" (p. 135). Ard claimed that attention to factors relevant to appropriate classroom discourse is not common among ITAs. Moreover, he suggested that since many foreign students who did their undergraduate education in their L1 are more competent in written discourse, the transfer from written discourse as a genre to spoken discourse as a different genre might be a difficult process. Ard proposed that, at the pedagogical level, negotiated and comprehensible input might help ITAs acquire realistic college discourse.

Yule and Hoffman (1990) looked at the ITA issue from an assessment-oriented perspective. They investigated the power of Test of English as a Foreign Language (TOEFL) and the verbal section of Graduate Record Exam (GRE) scores in predicating ITA candidates' success in being officially admitted as TAs. In the course of two years, Yule and Hoffman used test records and final evaluations of 233 graduate students who were awarded teaching assistantships. The final evaluations were made based on the student's performance in the ITA training course that was mandatory for all TAship international awardees. The researchers found that there is a significant correlation between TOEFL and verbal GRE and positive evaluations at the end of the ITA training course. They recommend a cutoff score of around 560 for the paper-based TOEFL for awarding assistantships that involve instructional duties. However, Yule and

Hoffman did not justify why the absence of a speaking module in both tests would not adversely affect their predicting power considering the crucial role that spoken discourse plays in fulfilling teaching duties.

Hoekje and Williams (1992) analyzed ITA education and assessment within the communicative competence framework (see Canale & Swain, 1980; Hymes, 1972). They argued that ITA proficiency involves both “appropriateness and correctness” (p. 246). Coupled with calling for a thorough needs analysis of ITAs, Hoekje and Williams stressed that the ultimate goal of ITA education is “effective language usage while performing the role of TA” (p. 247). In their opinion, this goal may be defined in terms of linguistic competence, pragmatic competence, and teaching abilities. Concerning testing ITAs, they raised questions about the validity of common proficiency tests mainly because of the tasks used in them. Hoekje and Williams emphasized that ITA assessment includes what is beyond linguistic issues, namely culture and behaviors and norms of American classrooms.

The concept of authenticity in language testing has been a major concern for language testers and test developers for quite a long time now. Screening ITAs’ proficiency, which arguably shares numerous facets of general L2/FL oral proficiency assessment, is not an exception when it comes to the issue of authenticity. Relying on Bachman’s authenticity framework, Hoekje and Linnell (1994) compared three different instruments to evaluate spoken English ability of ITA candidates. These tests included SPEAK (Spoken Proficiency English Assessment Kit) test, the ACTFL OPI (American Council on the Teaching of Foreign Languages Oral Proficiency Interview), and IP (Interactive Performance) test which was a teaching performance test exclusively designed for ITAs at the university where the data were collected. The authors claimed that the communicative situation involving ITAs is “interactive” and “complex” in nature (pp. 106-107). They compared the selected instruments based on Bachman’s (1990) test method facets: environment, rubric, input, response, and interaction between input and response (p. 112). The results showed that, in task authenticity, IP is the closest to the tasks ITAs perform in related contexts. Hoekje and Linnell (1994) drew this conclusion based on the premise that, in IP, test takers were able to manipulate and control the task, as well as “negotiate turn taking among the speakers” (p. 113). Furthermore, the extensive amount of discourse distinguishes a performance test such as IP from SPEAK or OPI. The type of interaction in IP was found to be the most authentic against the type of interaction in SPEAK (i.e. audiotaped, no interlocutor present) and OPI (i.e. face-to-face with one interviewer/interlocutor). However, Hoekje and Linnell suggested that tests like IP are unable to assess certain sociolinguistic features (e.g. “the speaker's sociolinguistic appropriateness with undergraduate students”, p. 121).

In a Canadian context, Saif (2002) investigated the washback effect of implementing an ESP test, exclusively designed to assess the speaking abilities of ITAs, in an ITA training course. The author proposed a model that is based on ITAs’ “needs”, “means” of assessing those needs, and its “consequences” for an ITA training program (p. 4). The results obtained through interviews and class observations suggested a positive washback effect on the “content of the teaching” (p. 27). Saif also reported improvement in the learning ability of the ITAs in her study. Another important finding of Saif’s study was the reactions of raters to the test itself. The majority of the raters considered the test fruitful, authentic, and geared to the nature of ITA proficiency. Also, almost all of them believed that the test would play a motivating role for ITAs to improve their speaking ability. Finally, the author argues that there is a “complexity” involved in the washback effect of an ITA ESP test “embracing both test effects and the educational changes resulting from such effects” (p. 30).

Farnsworth (2013) investigated the construct validity of the speaking module of one of the most popular test of language proficiency worldwide, that is, TOEFL iBT, for the purpose of screening ITAs. Contrary to the claims made by Hoekje and Linnell about the OPI and SPEAK (1994), Farnsworth found that TOEFL iBT is a good assessment tool to evaluate proficiency of ITAs. Using factor analysis on 85 test samples, he concluded that TOEFL iBT measured the same construct in ITA proficiency when compared to a test specifically designed for ITAs. Farnsworth argued that in terms of dependability and practicality, TEOFL iBT subscores in speaking could be reliably used for ITA certification purposes.

In another study focusing on TOEFL iBT and international graduate students’ L2 oral performance, Brooks and Swain (2014) compared speech samples from the speaking section of TOEFL iBT with the real-life performances in and outside the classroom setting. They used background questionnaires and semi-structured interviews to triangulate data from TOEFL iBT performances of 30 ESL graduate students in Canada. Brooks and Swain relied on syntactic and lexical complexity measures, as well as accuracy measures[1] and measures to assess the use of discourse features. Altogether, Brooks and Swain found mixed results that both do and do not support the argument validity for the speaking section of TOEFL iBT. Specifically, the researchers reported ‘patterns’ in the responses that pinpointed a decrease in syntactic complexity from TOEFL iBT to non-test situations (p. 568). Furthermore, unexpectedly, TOEFL performances turned out to be the least accurate. To conclude, Brooks and Swain argued that,

[1] Literature on the measures of complexity, accuracy, and fluency are reviewed in the upcoming section.

based on the premises of validity argument (Chapelle, 2012; Kane, 2012), “comparisons of speaking in the test and real-life academic contexts show both overlap and non-overlap of performances” and this was an indicator of “potential weak link in the interpretive argument chain” (p. 371).

2.3. Research on CAF constructs

Research on complexity, accuracy, and fluency is abundant in the literature on L2 acquisition and performance. CAF constructs (or measures) owe their popularity to the fact that, according to Pallotti (2009), they can “describe” learner’s performance in the form of “variables” that “assess variation” (p. 590). Also, Skehan (2009, p. 510) refers to CAF constructs as “useful measures” to analyze performance in L2. Accordingly, Larsen-Freeman (2006) stressed the centrality of variability in learner language. However, Pallotti warns that lack of variation described by CAF does not mean that results are not scientifically acceptable. Housen and Kuiken (2009, p. 461) describe these measures as:

> “CAF have been used both 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.”

Numerous studies have investigated CAF in SLA research to track the development of learner’s proficiency. The importance of the CAF triad has been to a level that, with the contribution of a number of world-renowned applied linguists, the journal of *Applied Linguistics* devoted a complete issue in 2009 to CAF. In the same vein, in 2012, Housen, Kuiken, and Vedder published a valuable anthology of current theoretical and empirical research on CAF. In the first chapter of their book, Housen et al. (2012) described the CAF framework as "the primary epiphenomena of the psycholinguistic processes and mechanisms underlying the acquisition, representation and processing of L2 system" (p. 2). As I briefly discussed in the previous chapter, the multilayered nature of CAF constructs has created a venue for continuous research that often leads to alterations and reformations of their previous definitions (For a review of studies incorporating CAF, see Plonsky & Kim, 2016).

2.3.1. Studies incorporating all CAF components and their interaction

CAF constructs are objective measures of language performance (Skehan & Foster, 1997). In one of the earliest studies involving such measures, Halleck (1995) compared OPI holistic proficiency measures (i.e. intermediate, advanced, and superior) to objective measures of complexity and accuracy (i.e. mean T-Unit length, mean error-free T-Unit length, and percent of error-free T-Units). Studying an EFL context in China, Halleck found that when it comes to complexity (referred to as syntactic maturity in Halleck's study), holistic ratings of proficiency levels would not cause a difference in proficiency. On the contrary, accuracy measures of mean error-free T-Unit length and percent of error-free T-Units made a difference between intermediate and superior learners, as well as between advanced and superior learners (p. 230). Halleck found no task influence (referred to as task variability in her study) regarding the studied objective measures (p. 231). However, this study did not use any of the objective measures of fluency. Overall, Halleck suggested that T-Unit is an adequate measure for the analysis of spoken data.

In search of a reliable and valid unit to measure accuracy and complexity in speaking, Foster, Tonkyn, and Wigglesworth (2000) reviewed and studied different "units of segmentation" in the literature (p. 371). They suggested that, in comparison with T-units, Analysis of Speech Unit (or AS-unit) is a more reliable measure to be used in analyzing oral data, since it also considers incomplete grammatical units that occur in natural speech (e.g. *Same here!*). Foster and her fellow researchers suggested that one had better exclude certain phenomena that constitute disfluencies, like false starts and repetitions, from AS-unit counts.

With reference to Robinson's (2003, 2005) *Cognition Hypothesis*, Michel, Kuiken, and Vedder (2007) empirically investigated this hypothesis in L2 oral performance. Cognition Hypothesis, as I also briefly explained in the previous chapter, stresses that "cognitively complex tasks trigger both greater accuracy and greater linguistic complexity", whereas the third component of the triad, fluency, "suffers from increased task complexity" (Michel et al., 2007, p. 242). Moreover, Cognition Hypothesis asserts that task type (dialogic vs. monologic) affects syntactic complexity differently, that is, dialogic and interactive tasks decrease syntactic complexity but monologic and non-interactive tasks trigger more syntactically complex language (see Robinson, 2011). In their study, Michel et al. used 44 Moroccan and Turkish learners of Dutch who performed monologic and dialogic tasks. Based on the results, more complex tasks elicited more accurate (number of errors per AS-unit) but less fluent (unpruned speech rate) language, although they did not influence syntactic complexity. Concerning task type, in comparison with monologic tasks, dialogic tasks stimulated more accurate and more fluent speech but less complex language. According to Michel et al., the results of their study did not "confirm the predictions of the Cognition Hypothesis with respect to measures of accuracy and linguistic [or overall] complexity" (p. 256). Finally, to secure the results of such studies, the researchers suggest that Robinson's hypothesis needs to be tested with L1 speakers as well.

Emphasizing the premise that *communicative adequacy* is the ultimate goal of many language learners, Révész, Ekiert, and Torgersen (2014) investigated the predictive value of various CAF measures in both L1 and L2 performance. They defined communicative adequacy as "the knowledge and employment of both linguistic and interactional resources in social contexts" (p. 2) which aligns largely to the principals of communicative competence. Their study compared an array of measures for each CAF construct with regard to task and proficiency level variables. The results indicated that an index of fluency (i.e. frequency of filled pauses) is the strongest predictor of fulfilling an oral task, and therefore, being communicatively adequate. In addition to filled pauses, false starts (another fluency index) were found to be a reliable predictor for advanced learners' adequacy. Similar to Halleck's (1995) finding, Révész et al. concluded that task type has no influence on CAF and adequacy.

Investigating the impact of task complexity on CAF constructs, Malicka (2014) found that accuracy improved in high proficiency learners with more complex tasks. The subjects were college-level tourism students from Catalan, Spain. Malicka claimed that as the cognitive load of the task increases, the production becomes more accurate. In addition, the results showed that low proficiency learners produced more syntactically complex language when the measure was number of words per AS-unit. Malicka reported a trade-off effect between two syntactic complexity measures of number of words per AS-unit and number of words per clause. Besides, the lexical complexity of the participants' production amplified with the increase in task complexity. With reference to fluency, the findings of this study revealed that tasks that were more complex caused no change in the speech rate (a fluency measure) of low proficiency learners, whereas high proficiency learners experienced a decrease in rate of speech.

In a longitudinal study, Vercellotti (2015) studied the growth of and competition between CAF constructs in oral L2 performance drawing on Skehan's (1998) *Limited Capacity Hypothesis* and Robinson's (2003) *Cognition Hypothesis*. The students enrolled in an intensive English program performed a monologic oral task twice during the course. Besides investigating lexical variety, Vercellotti used AS-unit to assess grammatical complexity, percentage of error-free clauses to evaluate accuracy, and mean length of pause to track fluency (p. 8). Overall, the researcher confirmed her hypothesis that "all measures" would grow upon receiving instruction (p. 9). The results also indicated that lexical variety was a predictor of fluency, grammatical complexity, and accuracy. More importantly though, Vercellotti found that "all measures showed growth over time" and "the within-individual correlations were positive" (p. 15). In other words, the learners in her study did not sacrifice any of the CAF components for the sake of being more complex, accurate, or fluent.

Concentrating on pictorial narrative prompts, de Jong and Vercellotti (2016) studied the effect of task type complexity on complexity, accuracy, fluency, and lexis when the proficiency level is not a variable. The findings revealed that different picture-based tasks steered no differences in performances in terms of complexity and accuracy, suggesting that task complexity was similar between tasks. However, fluency turned out to be different across some of the prompts. Lexical variety, on the other hand, was different for all the tasks in de Jong and Vercellotti's study.

2.3.2. Research on complexity

The triad starts with the least agreed-upon of the three: Complexity. Many of the studies published on complexity deal with either justifying previous definitions or suggesting a new (or revised) definition of the construct (see Pallotti, 2015). These are some of the definitions of complexity in the literature:

- "The extent to which the language produced in performing a task is elaborate and varied" (Ellis, 2003, p. 340);
- "[T]he capacity to use more advanced language" (Ellis, 2009, p. 475);
- "[L]anguage that is at the upper limit of the student's interlanguage system, which is not fully internalized or automatized by the learner" (Vercellotti, 2012, p. 14, based on Ellis & Barkhuizen, 2005).

Nevertheless, most of the CAF researchers have not limited their definition of the term (and consequently their operationalization of the construct) to simplistic and straightforward descriptions provided above. Housen and Kuiken (2009) distinguished *cognitive* complexity and *linguistic* complexity. The former is relative, and essentially, refers to difficulty or cognitive load associated with performing a task in L2, while the latter denotes an objective account of the learner's L2 system (see also DeKeyser, 2008, in Housen & Kuiken, 2009). From a performance-based standpoint, Pallotti (2009) argued that complexity is grammatical and can be described as *lexical* or *syntactic*. Pallotti (2009) stated that, for a linguistic form, being more complex does not mean that the form is necessarily more developed (p. 593). In another article on complexity, Pallotti (2015) suggested a "simple view" of this multifaceted construct and recommended that complexity could be scrutinized merely in terms of "structural, formal aspects of texts and linguistic systems" (p. 118). Pallotti (2015, p. 120) even went further to propose that structural complexity can be classified into two categories of *grammatical* complexity (i.e. syntax rules in a language) and *stylistic* complexity (i.e. idiosyncratic rhetorical choices made by individuals). Grammatical complexity, according to Pallotti (2015), includes *lexical*, *syntactic*, or *morphological* complexity. As I discussed in the previous chapter, the present study used a structural view of complexity in its design. Moreover, Bulté and Housen (2012, 2015) tried to demonstrate the compound and componential nature of complexity in their publications. Figure 1 dispalys their taxonomic representation of L2 complexity that depicts the multidimensionality of this construct:

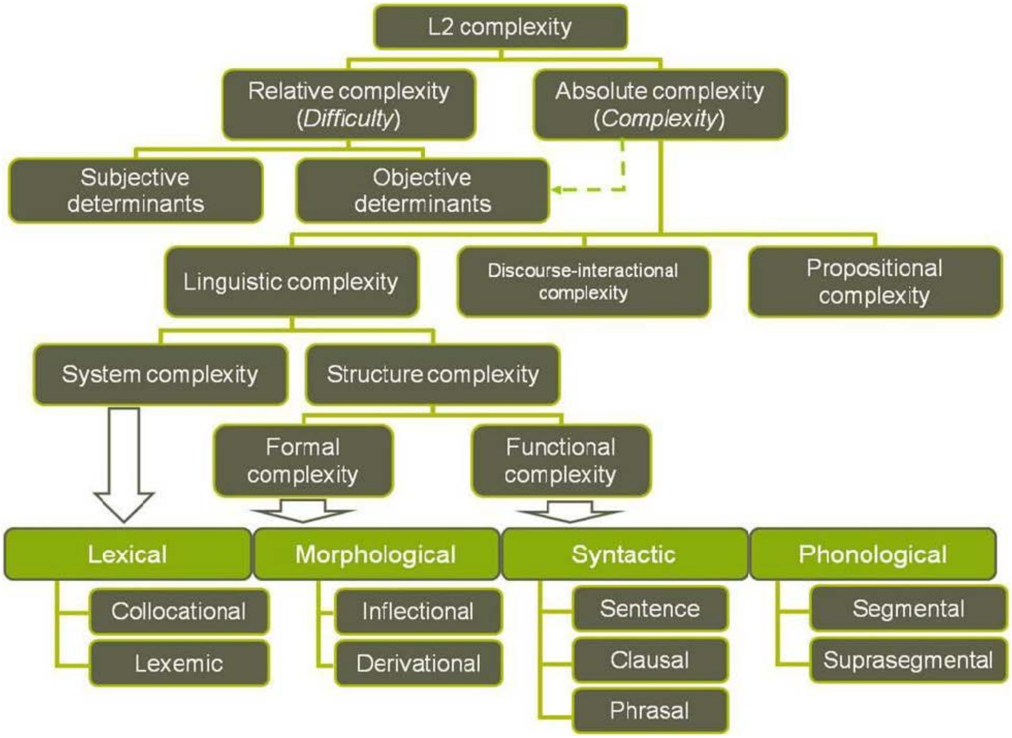


Figure 1. Bulté and Housen's (2012) taxonomy of L2 complexity. Reprinted from *Dimensions of L2 Performance and Proficiency* (p. 23), by A. Housen, F. Kuiken, I. Vedder, 2012, Amsterdam: John Benjamins. Reprinted with permission.

Most of the studies in SLA and L2 assessment incorporating complexity as a predictor in oral proficiency viewed this construct as lexical, syntactic, or both (see Norris & Ortega, 2009 for a review of the studies on complexity measures). For instance, Iwashita (2006) studied the common syntactic complexity measures to predict success in oral production of Japanese as a foreign language (JFL). Recorded narratives from 33 JFL learners at two different proficiency levels were used in this study. Following Halleck (1995), Iwashita utilized T-unit as the unit of production (or segmentation) to measure complexity of the oral data. The researcher explained that AS-units were not used "due to the complex procedure involved in coding" them (Iwashita, p. 158). Comparison of the two groups revealed significant differences, meaning that more advanced leaners used more units; therefore, their production was more complex. High proficient leaners used longer T-units with more independent clauses. Moreover, Iwashita reported that T-unit length had high correlation with task type. The number of clauses per T-unit was also found to be a valid predictor of JFL oral proficiency.

Rimmer (2006) stated that knowledge and mastery over syntax is "a significant factor in differentiating between score levels and characterizing overall proficiency" (p. 497). Reviewing measures on syntactic complexity, he challenged unit length as a "strong indicator" of complexity (p. 506). Rimmer argued that the reason that researchers tend to use unit length more often is the ease of reporting in quantified data that might draw an unrealistic picture of learner's proficiency, particularly if the performance is oral. Rimmer concluded by suggesting that, to explore a valid measure of syntactic complexity, corpus-based data can be efficiently incorporated, since "an attractive principle for an understanding of complexity is frequency" (p. 508).

A section of Inoue's (2016) study, which was concerned with syntactic complexity, identified measures that best fit various proficiency levels across different tasks. Utilizing AS-units, Inoue analyzed spoken narratives of Japanese EFL learners in three proficiency levels (beginner, intermediate, and advanced). Based on the results, coordination per AS-unit and AS-unit length did not significantly correlate with task type. For predicting proficiency levels, unlike what Norris and Ortega (2009) found on L2 writing, "for the spoken narrative performances", coordination index (i.e. "the amount of coordination") was not "the best predictor" (Inoue, 2016, p. 7). Finally, Inoue advised that, to test syntactic complexity measures, cautious piloting of tasks is mandatory due to "differing degrees of task-essentialness" (p. 13).

Vercellotti and Packer (2016) investigated the development of learners' sentence-level syntactic complexity in an English for academic purposes (EAP) setting. Considering clauses as the base for studying complexity, the researchers focused on six clauses (main or sentential, coordinate, adverbial, relative, complement-taking predicate or CTP, and nonfinite) to analyze 227 monologic speech samples. The participants' proficiency levels were low-intermediate, high-intermediate, and low-advanced. Based on the coded data, Vercellotti and Packer proposed a development order that is displayed in Figure 2. They reported that they did not find enough evidence to trace the development order of coordinates. Inclusively, adverbial clauses emerged earlier than other clause types and learners produced nonfinite clauses more frequently as their proficiency increased (p. 188).

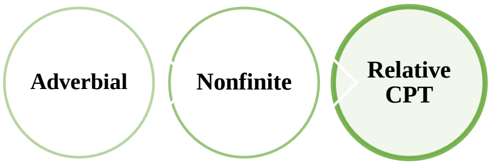


Figure 2. Vercellotti and Packer's (2016) proposed order of clause development

2.3.3. Research on accuracy

In SLA, according to Foster and Wigglesworth (2016, p. 98), "the route to acquisition is characterized by movement towards accuracy in performance." Many researchers who investigate CAF emphasize the importance of accuracy as a thorough indicator of success in communication both in written and spoken modalities. Pallotti (2009, p. 592) labeled accuracy as the "most internally coherent construct" of the CAF triad. Housen and Kuiken (2009) stated that accurate language connotes language that is deprived of errors and they defined errors as "[d]eviations from the norm" (p. 463). Nonetheless, the issue of appropriately defining norms does not allow an effortless recognition of what constitutes an error and what does not. Besides, as Pallotti (2009) and Wolfe-Quintero, Ingaki, and Kim (1998) argued, many accuracy measures try to compare L2 learner's performance with target-like norms. Although this may not yield satisfactory results at all stages of learning and/or acquisition development, Norris and Ortega (2012) believed that comparison with native speaker norms might be more justifiable in advanced levels.

Based on the arguments surrounding the limitations on working memory and *Trade-off Hypothesis* (see Skehan, 1998, 2014), certain task/post-task conditions trigger learners to produce language that is more accurate. For instance, familiar content, clear instructions, and straightforward structure, as well as transcribing the performance upon completion of the task can lead to high accuracy (Skehan, 2009).

Yuan and Ellis (2003) studied the impact of pre-task planning and online planning on CAF in monologic oral performance of L2 users. Despite the fact that they investigated all CAF components, here, I will only review the accuracy variable. Arguing that previous research has shown improvement only in fluency and complexity, the researchers found that online planning could also boost accuracy. Based on Levelt's (1989) speech processing model, Yuan and Ellis defined online (or on-line) planning as producing "careful" language along with "monitoring" speech at both "pre-production and post-production" stages (pp. 5-6). Planning, as Yuan and Ellis discovered, can provide the speakers with the opportunity to access their syntactic repertoire faster, and therefore, produce more accurately. The accuracy measures that were utilized in this study on undergraduate students in China included percentage of error-free clauses and percentage of correctly used verbs (i.e. accurate in terms of tense, aspect, agreement, etc.). The results of the measure count for three groups (no planning, pre-task planning, and online planning) indicated that the students who had been given the chance to perform online planning produced more accurate language (i.e. both more accurate verb forms and more accurate clauses) than the other two groups.

In a broad review of the studies on various measures on accuracy, Foster and Wigglesworth (2016) divided the existing measures into two main categories: local measures (i.e. measuring the use of a specific grammatical form) and global measures (i.e. analyzing the general accuracy level of the performance). As the authors argued, local measures are suitable to track the development of a particular form (e.g. morpheme inflections or verb tense). However, by referring to Ellis and Barkhuizen (2005), they warned that the SLA literature questions ‘concurrent’ development of most of the grammatical features, and this makes the validity of local measures disputed. On the contrary, global measures assess speech “in its entirety” based on “error rate” (Foster & Wigglesworth, 2016, p. 102). The error rate, according to Foster and Wigglesworth, can be calculated by:

- Using units of segmentation to divide the sample (e.g. 100 words per clause/T-unit/AS-unit)
- Calculating the distribution of error-free units (e.g. % of error-free clauses/T-units/AS-units)

The authors also advise the researchers about the fact that drawing syntactic boundaries is not an easy task and a consistent approach needs to be adopted. Moreover, Foster and

Wigglesworth challenged the processes involved in error identification and suggested that research should “base [the] analysis on syntactic units such as clauses, T-units, and AS-units, which can be more reliably identified” (p. 103). In addition, considering *error gravity*, they suggested a Weighted Clause Ration (WCR) as a new measure for accuracy that also takes into account magnitude of errors. Table 1 summarizes the suggested categorizations of errors and their corresponding scores for quantitative analysis:

Table 1. *Error categorization for Weighted Clause Ratio as a measure of accuracy (Foster & Wigglesworth, 2016, pp-106-107)*

Error Category	Score	Description
Completely accurate	1.00	The clause is fully accurate.
Level 1	0.80	Minor errors are detectable. Meaning is communicated.
Level 2	0.50	Serious errors are detectable. Meaning is somewhat lost.
Level 3	0.10	Numerous errors are detectable that seriously distort meaning.

WCR can be calculated through dividing the raw total score (i.e. sum of the scores for all four error categories) divided by the total number of clauses. Forster and Wigglesworth argued that their suggested measure taps both “morphosyntactic and semantic accuracy” in a reliable way (pp. 112-113). However, the researchers did not explain how raters of accuracy could harmonize their decision on whether an error definitively fits the description of the category to which it is assigned.

I previously reviewed a section of Inoue’s (2016) article on complexity. The second section of her study dealt with identifying valid measures to evaluate accuracy in oral performance. She researched narratives of two different tasks across three proficiency levels in an EFL context in Japan. Inoue compared correlations between tasks and three accuracy measures of percentage of error-free clauses, errors per AS-unit, and errors per 100 words (p. 10). Results of the statistical analysis revealed that errors per 100 words correlated more with the tasks. Furthermore, Inoue argued that different units of segmentation could produce varying results when it comes to assessing accuracy. Analyzing errors per 100 words, she found that the longer task (i.e. the one with more words) produced more errors. In contrast, there were longer AS-units in the shorter task that resulted in more errors when using this unit of segmentation.

2.3.4. Research on fluency

Fluency is the third component of the CAF triad, and similar to complexity, it has a multifaceted nature (Pallotti, 2009; Tavakoli, 2016). I discussed in the previous chapter that Lennon (1990, 2000) regarded proficiency in a ‘broad’ and in a ‘narrow’ sense. The broad definition of fluency views it as general proficiency or the ability to communicate efficiently in a language. However, the second definition may concern us more as applied linguistics researchers and/or language testers (or whatever SLA identity to which we associate ourselves). In the narrow sense, according to de Jong, Groenhout, Schoonen, and Hulstijn (2015), fluency is “described in terms of speedy and smooth delivery of speech without (filled) pauses, repetitions, and repairs” (p. 224). Some researchers who study fluency in L2 development and assessment believe that, in order to arrive at a reliable account of what constitutes fluent L2 speech, one should analyze data from L1 and L2 alongside each other (e.g. Segalowitz, 2010). de Jong and his colleagues divided the construct of fluency into three main categories of *cognitive*, *utterance*, and *perceived* fluency:

> “[C]ognitive fluency [is] the ability of the speaker to smoothly translate thoughts to speech. However, this ability cannot be measured directly.

> Therefore, researchers use measures of utterance fluency to gauge speech-planning difficulties that surface in utterances by counting the number of filled pauses, corrections, and repairs, and by measuring the duration of pauses. Yet another sense of fluency is perceived fluency, which pertains to the inference listeners (raters) make on the basis of the utterance about speakers' ability (about speakers' cognitive fluency)" (de Jong et al., 2015, p. 225).

de Jong et al. classified utterance fluency into *speed* fluency that accounts for the rate of the speech, *breakdown* fluency, that is silence and pausing, and *repair* fluency which deals with reformulations and hesitations. Figure 3 summarizes different types of fluency based on de Jong et al (2015) and Segalowitz (2010):

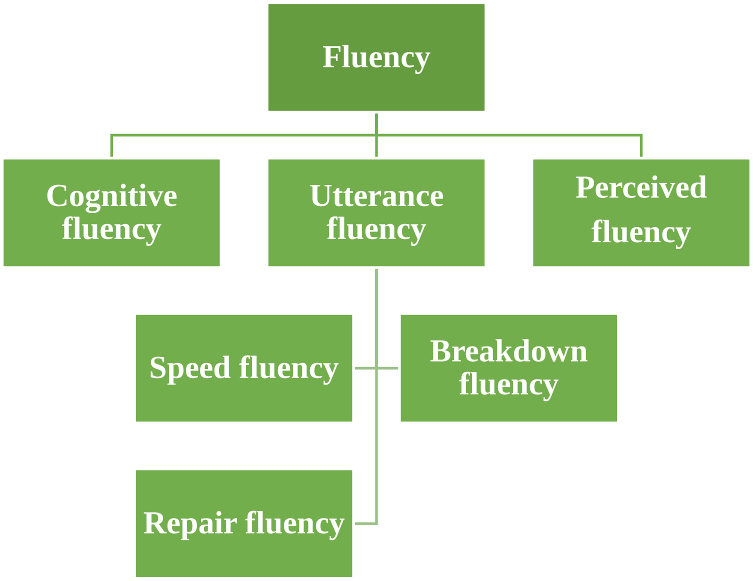


Figure 3. Typology of speech fluency (Based on de Jong et al., 2015; Segalowitz, 2010)

In a rather different framework, Skehan (2014) suggested that there is a difference between the nature of fluencies measuring speakers' *speed* and those focusing on the *flow* in their speech. In other words, disfluencies that interrupt the flow are to be distinguished from disfluencies that affect the speed (Tavakoli, Campbell, & McCormack, 2016). This distinction in fluency types is displayed in Figure 4:

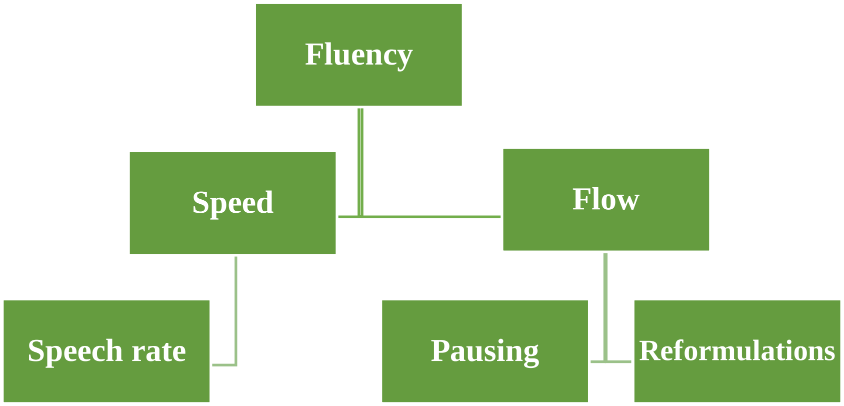


Figure 4. Skehan's (2014) framework of speech fluency

Several studies, some of which are reviewed below, have tried to find the most reliable measures to assess utterance fluency (i.e. the most frequently investigated type of fluency) in L2 development and performance. According to Tavakoli (2016) and Witton-Davies (2014), it seems that the following measures can be reliably used to analyze utterance fluency:

- Pausing
    - Length of pause
    - Frequency of pauses
    - location of the pause in the clause
- Speed
    - Speech rate

    - Articulation rate

- Phonation time (i.e. speaking time minus pauses)
- Mean length of run (i.e. mean number of syllables between pauses)[2]
- Repair measures (e.g. number of hesitations, reformulations, etc.)

Not a lot of studies have focused on either some or all of CAF components within the ITA framework. In one of such studies, Gorsuch (2011) emphasized the significance of fluency in helping ITAs succeed in their performance. She operationalized fluency in terms of “intact” vs. “split” pause groups where phrasal boundaries are syntactically observed as in the former and broken as in the latter (p. 3). In her study, Gorsuch compared the influence of repeated reading (RR) as the instructional input throughout an ITA training course on the improvement of ITA candidates’ fluency against a production-oriented control group. The results indicated that the RR input group used fewer split pause groups at the end of the instruction period. Gorsuch suggested that ITAs that are more fluent could plan at the discourse level to produce more intact pause groups. Overall, the study results advocated a blend of orientation and input-focused designs for ITA training courses (p. 5).

In a study on fluency measures, de Jong et al. (2015) investigated the role of L2 fluency measures in predicting L2 proficiency. Moreover, they studied whether “L2 fluency measures that are corrected for L1 fluency behavior” can predict L2 proficiency better than uncorrected measures (p. 226). Coupled with AS-units, certain measures of utterance fluency were used including mean duration of syllables (speed fluency), number of silent pauses, length of silent pauses, and number of nonlexical filled pauses (breakdown fluency), and number of repetitions

[2] Tavakoli (2016) classified phonation time and articulation rate as *composite* measures of fluency, that is, measures that track both speed and pausing in speech.

and number of corrections (repair fluency). de Jong and his colleagues found that only some of the fluency measures correlated with proficiency level. This was contrary to the findings of de Jong, Steinel, Florijn, Schoonen, and Hulstijn (2013) in which all measures of proficiency turned out to be related to L2 proficiency. In de Jong et al.'s (2015, p. 238) study, mean syllable duration was "strongly" correlated with proficiency in L2, whereas pause duration and proficiency did not significantly correlate. Additionally, the results revealed that "for the fluency measure syllable duration, a corrected score is more strongly related to a measure of L2 proficiency than is the original uncorrected L2 measure" (p. 237). Overall, concerning the role of L1, de Jong et al. (2015, p. 239) concluded that language tests can benefit from using "L1 behavior as a baseline" in their design, and learners can also benefit from modifying "their speaking style" either in L1 or L2.

Mirdamadi and de Jong (2015) compared the effect of syntactic complexity on utterance fluency in both L1 and L2. The researchers chose active and passive structures based on the hypothesis that, since passive voice is acquired later than active constructions, "producing a passive [would] be more difficult than producing an active… [i]t is likely that the passive structure is not as proceduralized and automatized as the active structure" (p. 108). College-level L1 speakers of Dutch reacted to cartoon images producing 40 sentences in English and 40 sentences in Dutch. Thus, Mirdamadi and de Jong employed articulation rate (a speed fluency measure) and number of hesitations (a hybrid measure of repair and breakdown fluencies). Analysis of the recorded passive and active sentences indicated that complex language (i.e. language including more passives) affected fluency by causing more hesitations in both L1 and L2 speech. It is worth noting that fluency in L1 was negatively influenced more than L2 fluency. Nonetheless, speech rate remained unaffected by syntactic complexity.

Tavakoli et al. (2016) studied the pedagogical aspects of L2 fluency development and its relationship with complexity and accuracy. The experimental group in the study, who were ESL students in an EAP course, experienced “awareness-raising activities” (e.g. listening to nonnative speakers’ speech samples) and fluency strategy training” (e.g. teaching the use of gap fillers) in the course of four weeks (p. 453). Analysis of the first minute of the participants’ monologic performance indicated that although the control group gained improvement in fluency, this improvement was much more significant for the experimental group. This improvement was mainly observed in fluency measures of “length of run, articulation and speech rates, and phonation time ratio” (p. 463). Furthermore, as Tavakoli et al. reported, “the development of breakdown fluency (i.e. silence and pausing) is slower and less sensitive to pedagogic intervention” (p. 464). Surprisingly, in connection with complexity and accuracy, it was only the control group who significantly improved in a specific measure of accuracy (i.e. percentage of error-free clauses). By contrast, the experimental group showed partial progress in these two constructs.

In a very recent article, Tavakoli (2016) scrutinized current measures of L2 fluency through comparing monologues and dialogues. She claimed the construct itself needs to be redefined, since research on fluency is overshadowed by “mixed results due to the lack of a systematic approach to measuring fluency” (p. 134). She argued that research on monologic tasks has been abundant due to the relative ease of analyzing data from such tasks, whereas the difficulty associated with measuring linguistic aspects of speech in dialogic tasks has limited their use in data collection. Following Tavakoli et al. (2016, see Chapter I), Tavakoli suggested that selecting fluency measures should be done carefully because of a possible overlap between some of these measures (see phonation time and mean length of run earlier in this section).

Finally, Tavakoli introduced two “dialogue only measures” in her article that can be used to describe fluency in a dialogic task (p. 139): number of turns and number of interruptions.

2.4. Research on compensatory strategies and communicative adequacy

It has long been argued that successful communication in a second language is the product of multiple competencies working together (Canale, 1983; Canale & Swain, 1980). However, sometimes, absence or lack of harmonized collaboration between such competencies causes communication breakdowns (Savignon, 1983). In other words, as Dörnyei and Scott (1997) put it, “the mismatch between L2 speakers’ linguistic resources and communicative intentions leads to a number of systematic language phenomena whose main function is to handle difficulties or breakdowns in communication” (p. 174). Such phenomena that merge with speaking abilities help L2 users cope with communication failures are mostly referred to as communication strategies. Nakatani (2006) defined communication strategies as “alternative plan[s]” that learners utilize to reach their communicative objectives “by means of whatever resources are available” to them (p. 174). Nakatani stressed that most of such communication strategies are useful to learners when they are involved in producing oral speech:

> “Oral communication strategies specifically focus on strategic behaviors
> that learners use when facing communication problems during interactional
> tasks (Nakatani, 2006, p. 152).”

Yaman and Özcan (2015) reviewed the classifications that exist in the literature on oral communication strategies. According to them, the studies on communication strategies are centered around two main views: The *interactional* view (Canale, 1983; Tarone, 1980) which accounts for “a mutual attempt by participants in a communicative situation to maintain communication" and the *psycholinguistic* view (Faerch & Kasper, 1983) which defines these strategies as "the individual's mental response to a problem" (Yaman & Özcan, 2015, p. 144). Both views, as Yaman and Özcan argued, narrow down their definition to different types of strategies that overlap in certain subcategories (see Table 2).

Table 2. *Typology of communication strategies (Reviewed by Yaman & Özcan, 2015)*

Communication Strategies				
Interactional view	Psycholinguistic view			
Approximation	Achievement strategies		Avoidance strategies	
Word coinage	Compensatory strategies	Retrieval strategies	Formal reduction	Functional reduction
Circumlocution	Code-switching			
Verbatim translation	L1 Transfer			
Code-switching/mixing	Cooperative strategies			
Asking for help				
Body language (or mime)	Nonlinguistic strategies			
Avoidance				

Poulisse (1987, 1997, in Dörnyei & Scott, 1997) grouped compensatory strategies into three categories of *substitution* strategies (i.e. replacing or modifying words in speech), *substitution-plus* strategies (i.e. foreignizing L1 for L2), and *reconceptualizing* strategies (i.e. altering the structure dramatically, e.g. circumlocution). Celce-Murcia, Dörnyei, and Thurrell (1995) proposed another typology of communication strategies. Based on Hymes' (1972) notions of communicative competence and Canale and Swain's (1980) model, Celce-Murcia et al. suggested that, in their model, the components of strategic competence include avoidance (or reduction) strategies, compensatory (or achievement) strategies, time-gaining strategies, selfmonitoring strategies, and interactional strategies. They described compensatory strategies as consisting of the following (Celce-Murcia et al., 1995, p. 28):

- “Circumlocution (e.g. *the thing you open bottles with* for *corkscrew*)
- Approximation (e.g. *fish* for *carp*)
- All-purpose words (e.g. *thing, thingamajig*)
- Non-linguistic means (*mime, pointing, gestures, drawing pictures*)
- Restructuring (e.g. *The bus was very... there were a lot of people on it*)
- Word-coinage (e.g. *vegetarianist*)
- Literal translation from LI
- Foreignizing (e.g. LI word with L2 pronunciation)
- Code switching to LI or L3
- Retrieval (e.g. *bro... bron... bronze*)”

In my study, as I briefly discussed in the first chapter, the focus was on the strategic competence, and therefore, compensatory strategies that are used to make up for ITAs’ linguistic deficiencies. Despite the comprehensive classifications that are available to researchers to use in the analysis of compensatory strategy use in spoken data, selection of strategies should be done with care when the subjects are ITAs. For example, certain utterance fluency measures consider retrieval or restructuring as disfluencies. In other words, a performance that comprises numerous retrievals cannot be appraised as an indicator of success when breakdown fluency is calculated for an ITA performance. I will elaborate more on this issue in the next chapter.

To my knowledge, the number of studies that investigated the use of compensatory strategies in ITA assessment and performance is very scarce. Halleck and Moder’s (1995) study is one of the few that examined and compared the role of ITAs' linguistic competence and compensatory strategy use in a performance test. The rating rubric that the researchers used in their study included categories on both language abilities and teaching skills. The graduate students who took part in the study "presented a 5-minute lecture in their field to a panel of two trained raters" labeled as the ITA Test (p. 740). Based on the rubric, participants were placed into three groups of passed, provisionally passed, and failed, with students in the 'passed' group "indicating only minor difficulties that do not interfere with comprehensibility" and showing readiness "for classroom teaching with no further training" (p. 742). Arguing that success in the ITA Test is the product of both linguistic and strategic competencies, Halleck and Moder concluded that after taking the ITA training course, the provisionally passed and failed groups showed low to moderated improvement in terms of linguistic abilities. However, their gains in the strategic skills (referred to as teaching skills in the study) was different between the two groups, that is, the provisionally passed ITAs eventually passed the test by surpassing scores in the strategic competence. To sum up, Halleck and Moder claimed that there is a threshold level of proficiency for ITA candidates to benefit from instruction of compensatory strategies, below which training in such strategies might be futile.

Jenkins and Parra (2003) examined the effect of nonverbal behavior and paralinguistic features as complementary tools for ESL learners when taking an oral proficiency test. Eight ITAs (4 Chinese L1s and 4 Spanish L1s) took part in a speaking test (interview format) exclusively developed for testing the ITAs at the institution from which data were collected. Through discourse analyses of the videotaped interview performances, Jenkins and Parra found that raters were sensitive to interviewees' nonverbal behavior during the interaction. That is, raters interpreted "active nonverbal behavior" and "appropriate paralinguistic features" as indicators of ITAs having interactional competence (p. 90). The researchers also argued that "borderline" students benefit more from employing successful nonverbal behavior alongside linguistic competence (p. 103). Nonetheless, in line with Halleck and Moder (1995), Jenkins and Parra suggested that very low proficiency students might not be successful by only relying on paralanguage as a compensatory strategy.

A part of the research on communication strategies deals with investigating strategy use among certain learner groups using inventories that already exist (e.g. Oral Communication Strategy Inventory by Nakatani, 2006). Yaman and Özcan (2015), for instance, utilized Nakatani's inventory to study the oral communication strategy use of Turkish EFL learners with regard to their language proficiency. They found that strategies for meaning negotiation, affective strategies, and compensatory strategies were the most frequently used strategies, while message abandonment and planning strategies were not popular among EFL students in Turkey (p. 153). Proficiency level (i.e. intermediate and advanced) was a significant factor in strategy use when it came to the least frequently used strategies.

In another EFL context, Rabab'ah (2016) studied the impact of explicit training in communication strategies on the oral ability of 124 Jordanian (Arabic L1) learners. Upon what Dörnyei and Scott (1997) recommended, the strategies taught included asking for help/repetition/clarification, requesting confirmation, self-repair, guessing, and circumlocution. Rabab'ah compared both Overall and Speaking scores from two administrations of the IELTS (before and after the instruction) and found that the experimental group gained significantly better overall scores in the second administration. Moreover, comparison of the scores of the speaking section of the test also proved that explicit instruction of the strategies yielded better results. The learners in the experimental group used more strategies in their oral performance than the learners in the control group. Finally, analysis of the type of strategies used indicated that, after instruction, the learners used more of what Faerch and Kasper (1983) labeled as achievement strategies, namely asking for help from the interlocutor and self-repair. However, asking for clarification, requesting confirmation, and guessing yielded no statistically significant differences between the experimental and control groups.

2.5. The missing link

Housen et al. (2012) stressed the viability of CAF constructs in describing L2 performance. Numerous studies, some of which I reviewed in this chapter, have already investigated various aspects of oral proficiency using CAF. Such publications include studies involving different L2 learner/user groups, such as immigrants (e.g. Mirdamadi & de Jong, 2015), L2 students in university-sponsored language programs (e.g. Révész et al., 2014), learners in intensive language programs (Vercellotti, 2015), language for specific purposes students (Tavakoli, 2016). Alas, as I discussed in chapter I, the research has failed to investigate the role of CAF constructs in describing the linguistic performance of ITAs. ITA proficiency is a unique, and arguably, independent construct from normal L2 oral ability. However, researchers are yet to come up with an agreeable operationalization of ITA proficiency due to its multilayered nature and varying subconstructs involved (Ard, 1987; Gorsuch, 2011; Halleck & Moder, 1995; Saif, 2002). Perhaps using CAF constructs and their measures could help describe the linguistic side of ITA proficiency (Mirshahidi & Saeli, 2016). Furthermore, a successful ITA performance is achieved through not only mastery over linguistic knowledge (i.e. competencies that can be measured by CAF), but also through retaining competencies that go beyond the language level to complement the language and also compensate for its shortcomings. Thus, research needs to focus on investigating both CAF and use of communication strategies (in this case, compensatory strategies).

In the next chapter, I will explain the methods and procedures that I employed to explore the role that CAF constructs and their measures as well as compensatory strategies play in predicting what constitutes success in an ITA performance test.

CHAPTER III.

METHODOLOGY

3.1. Overview

The previous chapter reviewed the research on variables of this study and pinpointed the gaps that exist in the literature concerning the potential role that complexity, accuracy, and fluency (CAF) constructs and compensatory strategies might play in predicting international teaching assistant (ITA) candidates' performance in an oral screening test. In this chapter, I will explain the methodology that I used to answer the research questions of the study, and subsequently, fill the gap in the research. The following methodology was devised to answer these research questions that I posed in the first chapter:

RQ1: Do CAF constructs and their related measures predict ITA candidates' performance in an ITA test? If so, what is the extent to which CAF constructs differentiate between ITA candidates based on their test performance?

RQ2: Is there a relationship between CAF constructs? If yes, how do CAF constructs influence each other when the task condition is not a variable?

RQ3: Do compensatory strategies predict ITA candidates' performance in an ITA test? If so, to what extent do compensatory strategies differentiate between ITA candidates based on their test performance?

RQ4: Upon comparing CAF measures and compensatory strategies, which one is a stronger predictor of ITA candidates' performance in the ITA test?

3.2. Participants

The participants of this study were 21 (Female = 9; Male = 12) international graduate students whose L1 was not English, studying at a southcentral university in the United States. These ITA candidates were in their mid-20s to early-30s. The language background of the participants included Mandarin Chinese, Korean, Turkish, Persian (or Farsi), and languages of India, such as Hindi, Telugu, and Punjabi. All the participants were already admitted as full-time students by the graduate college of the school at either the Master's or the Doctoral level. They[3] were enrolled in different graduate programs, including the following:

- Industrial Engineering (n = 7)
- Hotel and Restaurant Administration (n = 4)
- Civil Engineering (n = 3)
- Electrical and Computer Sciences (n = 3)
- Economics (n = 2)
- Microbiology (n = 1)
- Curriculum Studies (n = 1)

In addition to providing the graduate college with the required language proficiency score (i.e. TOEFL iBT score of 79 or IELTS Academic score of 6.5) for admission, these students were supposed to meet a minimum requirement of English speaking proficiency to be eligible to

[3] N = 21

apply for a TA position within their departments. I will explain the dynamics involved in the following section.

3.3. Context of the study

During the application process or after being admitted to the graduate school, international students could apply for financial assistance in the form of teaching assistantships (TAship). However, in the university where the data were collected, meeting the minimum overall proficiency score was not enough of a requirement to apply for the TAship. ITA candidates (i.e. nonnative speaker applicants for the TAship) had to provide the graduate college with an acceptable speaking score in English in order to be considered for the position. That is, those who had obtained a score of 26 or above on the Speaking module of the TOEFL iBT (or equivalent) were exempt from an oral performance test (henceforth, the ITA Test) administered by the ITA Program of the university. [4] Applicants who had reported a speaking score of 22 to 24, however, were required to take the ITA Test (see the next section for a detailed account of the test). Finally, students whose score on the Speaking section of TOEFL iBT was below 22 were not eligible to apply for the TA position (see Table 3).

Table 3. *Decisions on international graduate students' applications for TAship (Based on their score of the Speaking Module of TOEFL iBT)*

Speaking score	26 to 30*	22 to 24**	21 or below
ITA decision	Cleared for teaching	Must take the ITA Test	Not qualified at all

* 30 is the maximum score on the Speaking Module of TOEFL iBT.
** A score of 25 is not given in TOEFL iBT

[4] A faculty member of the TESL Program at the university's English Department administered the ITA Program. Nevertheless, the graduate college was responsible for monitoring activities of the ITA Program and administration of the ITA Test. In the years before I conducted this research, the university's assessment and testing office was in charge of collecting the scores and reporting them to the graduate college.

All the students who participated in this study belonged to the group who were required to take the performance test for the first time prior to be able to be certified to teach. In general, ITAs are basically expected to be upper-intermediate to advanced ESL speakers to be able to teach undergraduate-level classes (as the instructor of record or as the TA to the instructor) or manage lab sessions (Gorsuch, 2011).

3.4. Instruments and variables

This study was conducted using the videotaped teaching presentations of ITA candidates who, for the first time, took part in a mandatory performance test to be cleared to serve as TAs in their respective departments. Two raters using a holistic rating rubric rated each test performance, and then, I analyzed these rated performances based on the CAF framework and the use of compensatory strategies. Below, I will explain each of these instruments and variables in detail.

3.4.1. The ITA Test

Participants of this study presented in an in-house assessment, called the ITA Test, which was a performance test exclusively developed for screening ITA candidates. The ITA Test has been used in the university for over twenty years now to facilitate the recruitment of ITAs. The test was adapted from Smith, Meyers, and Burkhalter (1992), and it was used in Halleck and Moder's (1995) study, which I reviewed in the previous chapter. As Halleck and Moder described, the ITA Test "requires ITAs to teach a minilesson and respond to questions in a classroom setting" (p. 737). Adhering to Bachman's (1990, 1991) views on communicative language testing, Hoekje and Linnell (1994) defined authenticity in testing as "a quality of the relationship between features of the test and features of the nontest target-use context" (p. 104). Accordingly, the ITA Test was an attempt by the university's ITA Program to develop and administer an authentic, English for Specific Purposes (ESP) test that targeted the particular needs and proficiencies required for TAs to teach an undergraduate-level class or lab.

Simply put, the test was a five-minute oral performance during which the test takers *taught* a topic of their choice associated with their prospective teaching content in the form of a lecture-like *presentation*. In order to authenticate the test situation (i.e. a situation similar to a real-life classroom atmosphere) what followed the presentation part was a brief question-and-answer (Q&A) period during which the audience could ask content-related questions. The Q&A period usually took about 2-3 minutes, and the audience including the raters and the undergraduate representatives asked 1 to 3 questions depending on the topic of the presentation. In other words, the test comprised one monologic and one dialogic task. During the monologic performance, interruption rarely occurred. In the Q&A period, though, turn-taking situations and use of strategies relevant in one-on-one interactions happened on a frequent basis. Figure 5 shows the basic structure and sequencing of the ITA Test. Table 4 summarizes the key characteristics of the ITA Test based on Bachman’s (1990, 1991) test method facets.

Table 4. *The ITA Test characteristics based on Bachman's (1990, 1991) test method facets (Layout inspired by Hoekje & Linnell, 1994, p. 114)*

Test method facets	The ITA Test
Test organization	Fixed timing; optional content
Input	Real-life linguistic input in the form of questions from audience members (Q&A)
Expected response	Extensive discourse (5 minutes) for monologic task; limited but relevant response in Q&A; discourse revolving around the topic of the lesson
Relationship between test input and test taker’s response	Non-reciprocal during monologic task; reciprocal through Q&A with turn-taking and meaning negotiation

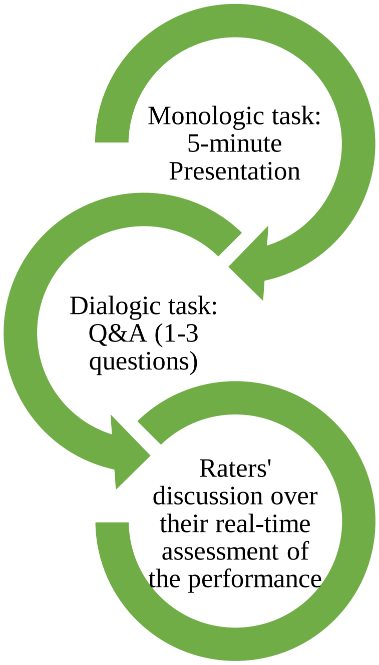

Monologic task: 5-minute Presentation
Dialogic task: Q&A (1-3 questions)
Raters' discussion over their real-time assessment of the performance

Figure 5. The ITA Test structure and sequencing

3.4.2. The ITA Test raters

The ITA Test performances were rated holistically by two raters in each test room, which was a regular classroom equipped with smart technology, such as video projector and computer. The raters were often faculty members and/or doctoral students from the university's TESL program who were already trained by the ITA Program's director. The training included observing test administrations and attending briefing sessions with the director. In case an alreay-trained rater was new to the ITA Test, they would be paired with a more experienced rater. In addition to the trained raters with TESL background, two volunteer undergraduate students were present in the test room as well to guarantee the situational and interactional authenticity of the test and help raters evaluate the performances; the first step for the raters to discuss the performances was to ask the undergraduate volunteers about their overall judgment of the presentations. The undergraduate volunteers' training process was much less vigorous than that of the main raters in that they were usually introduced to the task a day before or on the day of the test.

3.4.3. The rating process and the rubric

The raters were engaged in real-time assessment of the performances, that is, the rating was done as each ITA candidate was presenting their topic. Nonetheless, since there was more than one rater involved, the raters shared and discussed their evaluations to reach a final agreement after a certain number of presentations (e.g. after every five presentations). Each case was discussed separately. The discussion session usually started by the raters asking the undergraduate volunteers to express their overall opinion about the performance. A frequent question that the undergraduate students were asked was whether they would stay enrolled in or drop the class taught by the ITA candidate under scrutiny. As the next step, the raters would engage in a discussion over the scores they assigned to the candidate based on the ITA Test rating rubric. The discussion would be necessary when the assigned scores (and therefore the result categories that I will discuss in the next paragraphs) were drastically different from each other. In other words, it was likely for the raters to change their score in a certain subset of the test based on their discussion with the other rater over the performance. Upon reaching an agreement, the panel decided whether to *pass, provisionally pass,* or *fail* the candidate.

The instrument that was used by the raters to evaluate the presentations was a holistic rating rubric utilized by the ITA program for the past few years (see Appendix A). Despite being holistic, the rubric comprised three general subsets: *Linguistic Competence* (including “pronunciation, grammar fluency, and comprehensibility”), *Interactional Competence* (including “aural comprehension, the ability to respond appropriately and effectively to questions, and appropriate audience awareness)”, and *Strategic Competence* (including “organization of material, appropriate development of content, and effective use of strategies to compensate for linguistic weaknesses)”. Out of 30 maximum points, the distribution of the scores was: Linguistic Competence (15 points), Interactional Competence (10 points), and Strategic Competence (5 points). Lastly, the interpretation of the ratings encompassed three major decisions that are summarized in Table 5. The scores were multiplied by 10 when reported to the students (i.e. Maximum score of 300). This was originally done so that the scores would look similar to scores obtained on the Test of Spoken English, which had a maximum of 300 points (G. B. Halleck, personal communication, October 12, 2016).

Table 5. *Interpretation of the ITA Test scores*

The ITA Test score	250 to 300	240 to 249	230 or less
Interpretation	Pass	Provisional Pass	Fail

Passed ITAs were cleared to teach as TAs, whereas provisionally passed ITA candidates were supposed to take the two-credit ITA training course to be able to retake the test and keep their TAship. Finally, failed candidates could not be recruited as teaching TAs (not even provisionally like the second group) by their departments. However, after completing the basic ITA Training Course that focused more on linguistic issues (e.g. pronunciation), they could reapply to take the ITA Test.

It is worth mentioning that the rubric was a revised version of a rating guideline that was originally adapted from Smith et al. (1992). Therefore, it would be logical to claim that the ITA Test was a ‘semi-in-house’ assessment because of being partially adapted from another test. Finally, it is of particular note that the three general subsets of the rubric served as assistive tools for the raters in their discussion; therefore, after all, the decision was made holistically for each test taker.

3.4.4. Operationalization of the variables

This study included two main groups of independent variables, namely CAF constructs and compensatory strategies. Each group, however, comprises various measures. In the following sections, I will define and operationalize each variable based on the characteristics of the construct of ITA proficiency, as the dependent variable of this study.

3.4.4.1. Measuring complexity

As I introduced in Chapters I and II, complexity is generally defined as richness and elaborateness of the language (Ellis, 2003; House, Kuiken, & Vedder, 2012). To review, there are three main perspectives of complexity in the literature, including structural, developmental, and cognitive (see Pallotti, 2015). In this study, I approached complexity from a structural perspective. To be specific, I viewed complexity as the analysis of the number of components (or elements) in the produced speech. The other two views did not necessarily suit the purposes of this research for a couple of reasons: First, as Pallotti (2009) and Housen et al. (2012) argue, both cognitive and developmental notions of complexity are dynamic, because they are different across L2 learners. Furthermore, from these two perspectives, complexity can only be measured subjectively. Secondly, my research objectives were to predict and explain the proficiency of ITA candidates based on the results of a single administration of a performance test. Thus, a developmental perspective of complexity was of no logistic value here. In addition, cognitive complexity (or difficulty), because of its psycholinguistic account of the difficulties associated with processing and producing ‘certain’ linguistic forms, was not in line with the objectives of this project.

In addition to approaching complexity from a structural perspective, I relied on another categorization of complexity based on which grammatical complexity is divided into lexical and syntactic complexities (Housen & Kuiken, 2009). I explained in Chapter I that I only investigated syntactic complexity that connotes the number of grammatical components in a chosen unit of segmentation and the relationship between these components. Vercellotti (2015) argues that topic in L2 oral tasks yields varying results regarding lexical variety (a measure of lexical complexity). She emphasizes that certain topics might “encourage” lexical variation, while some trigger “inconsistent topic effects” (p. 15). Similarly, the other measures of lexical complexity, including lexical diversity and lexical density, are considered questionable by some researchers (e.g. Pallotti, 2015). Lexical density, for instance, is the ratio of lexical items (or non-function words) to the rest of the text. According to Pallotti (2015), “it is not clear whether a higher rate of lexical words should denote more or less complexity” (p. 126). Also, Bulté (2007) argues that lexical complexity measures based on type/token ratios are not valid. Overall, there are a lot of problems associated with choosing an appropriate measure of lexical complexity, and there is not enough agreement on the research-worthiness of such measures (McCarthy & Jarvis, 2010). Therefore, based on the fact that presentation/teaching topics varied across test takers in the ITA Test, and also based on the uncertainties in the literature concerning lexical complexity measures, I decided to focus only on the syntactic complexity of the ITA candidates’ oral performance.

The first step to measure syntactic complexity might be deciding on a unit of segmentation to be utilized consistently (Foster, Tonkyn, & Wigglesworth, 2000). Minimal terminable unit or T-unit (Hunt, 1965, in Norris & Ortega, 2009) and Analysis of Speech Unit or AS-unit (Foster et al., 2000) are the two most frequently used units in the studies on CAF. I argued earlier in this chapter that the ITA Test was developed to comprise one monologic and one dialogic task. However, the main section of the test was for the ITA candidate to teach a topic, that is, it was monologic in nature. In Brooks and Swain’s (2014) words, monologic tasks generally elicit “more formal discourse” (p. 356). Previous research on L2 writing development has shown that T-unit is a reliable unit of language for measuring syntactic complexity (Norris & Ortega, 2009; Wolfe-Quintero, Inagaki, & Kim, 1998). It goes without saying that writing requires more attention to formal discourse and rigid grammatical rules that are occasionally modified or ignored in oral speech. In the same vein, some studies stress the functionality of T-units in analysis of L2 spoken data, particularly for non-beginner adult L2 users (e.g. Halleck, 1995; Iwashita, 2006). Therefore, one can assume that T-unit is applicable to the analysis of spoken data when the task is monologic and the discourse is formal.

T-unit may be defined as “an independent clause and all of its dependent clauses” (Iwashita, 2006, p. 157). To clarify, in this study, a phrase such as Example (i)[5] was analyzed as one T-unit consisting of one independent clause (IC) and one dependent clause (DC). However, Example (ii), which is a compound sentence (for explanation of English clauses and sentences see Jacobs, 1995), includes three T-units, each comprising one IC.

(i) “*These two satellite portions are the railway stations where passengers can come in and go over the rails*.” → 1 T-unit, 1 IC, 1 DC

[5] All the examples in this section are excerpts from the transcribed ITA performances that were videotaped for this research project.

(ii) "*It can -- this is -- this can be used for the passengers and it can also used to transport the goods and it can also used to transport the automobiles*."[6]

→ 3 T-units, 3 ICs, 0 DC

General measures of complexity have gained more popularity as better predictors and descriptors of language performance, particularly when the development or production of a specific syntactic form is not the goal of the study (Housen & Kuiken, 2009). Many studies have used different measures of length and subordination to assess syntactic complexity (see Bulté & Housen, 2012, 2015). In this study, I used and compared two general measures of syntactic complexity: Mean length of T-unit (MLT) as a measure of length and proportion of clauses to T-units (Cs/T-unit) as a measure of subordination. Of particular note is that Iwashita (2006) labels number of clauses per T-unit as a 'general measure of complexity', whereas Révész, Ekiert, and Torgersen (2014) refer to a similar measure (i.e. proportion of clauses to AS-units) as a specific index of subordination.

3.4.4.2. Measuring accuracy

There are fewer complications involved in measuring accuracy in L2 oral production, since there is more agreement on the properties of this construct (Pallotti, 2009). Research on accuracy measures has indicated that global (or general) measures of accuracy are highly sensitive to spoken data (Skehan & Foster, 1999). Moreover, according to Foster & Wigglesworth (2016, p. 102), global measures of accuracy analyze speech "in its entirety," and not only particular form(s), which best suits the purpose of the present study (see also Ellis & Barkhuizen, 2005). Thus, I analyzed the accuracy of the oral performances of the candidates who

[6] In Example (ii), the first two disfluencies do not belong to the following T-unit because they are false starts (marked by double dash). Fluency measures will be entertained in section 3.4.4.3.

took the ITA Test through a global measure, namely percentage of error free T-units (see Example iii). As a reliable global measure, proportion of error free units, such as per 100 words, clauses, T-units, and AS-units, has been widely used in the studies on L2 accuracy (Foster & Skehan, 1999; Halleck, 1995; Révész et al., 2014; Tavakoli, Campbell, & McCormack, 2016; Vercellotti, 2015; Yuan & Ellis, 2003). Also, some studies have emphasized its success in differentiating between proficiency levels (e.g. Halleck, 1995).

(iii) *"These molecules are chemical words that the bacteria use to communicate with each other."* → An error-free T-unit

During the transcription process, I figured that, in some of the performances, the unintelligible words (i.e. words that were unable to be transcribed due to their unintelligibility despite acceptable audio quality) hindered comprehensibility of the sentence. I mentioned in the first chapter that, according to Tonkyn (2012), accuracy and comprehensibility cannot be treated separately due to the conflicting results in the literature. Moreover, unintelligible word were not treated as evidence of dysfluency (Tavakoli, 2016) since, in the oral data for this study, they did not slow down or pause the flow of speech as it was produced by the presenter. As a result, I decided to introduce the number of unintelligible words as a general measure of accuracy. To control for listener familiarity and bias (see Lindemann, 2017; Munro & Derwing, 1995, 2006), the audio files of the performances were also played for two volunteer native speakers of English who did not have prior familiarity with the accents. As the last step, I calculated the mean number of unintelligible words per 100 words. Example (iv) contains two unintelligible words (marked by double parentheses) that both transcribers and the two volunteers failed to understand despite the high quality of the audio files extracted from the videos.

(iv) *"They just count three operations (( )) which are in the (( )) and the doctors can even operate on the spot."*

To validate the number of unintelligible words per 100 words as a general measure of accuracy, a between-groups ANOVA was performed. The hypothesis was that this suggested measure would differentiate between the three ITA candidate groups, that is, Passed (*M* = .78, *SD* = .54), Provisionally passed (*M* = 1.44, *SD* = .62), and Failed (*M* = 2.93, *SD* = 2.01). At 95% confidence interval, the independent between groups ANOVA indicated a statistically significant effect, *F*(2, 18) = 4.57, *p = .028, η2* = .379. Thus, 37.9% of the variance in the number of unintelligible words was accounted for by the categories ITA candidates were assigned to based on the ITA Test results.

Finally, concerning the coding of the errors, the nonconformities in both grammar and lexicon were treated as errors (Lennon, 1995; Tonkyn, 2007, 2012). While grammatical accuracy includes appropriateness and conformity with grammatical norms and syntactic rules, lexical accuracy refers to the extent to which the words in the produced speech are chosen appropriately and within the accepted norms of the communicative context. Each of the following examples contain a grammatical error (GE); there is a subject-verb agreement error in the first sentence and a word form error in the second sentence.

(v) *"So, supply curve is the quantity of pizza that producer plan*$_{\text{(GE)}}$ *to sell at each price."*

(vi) *"Today, in discuss*$_{\text{(GE)}}$*, I would like to talk about how to calculate the consumer surplus."*

However, Example (vii) contains two lexical errors (LE), including using a word that does not exist in English (i.e. *belongness* instead of belonging) and a word that does not fit the global context of the speech (i.e. *members* instead of managers).

(vii) *"...I mean we use belongness*(LE) *as a*(GE) *members*(LE) *to motivate our employees."*

3.4.4.3. Measuring fluency

In its narrow sense, fluency refers to an objective constituent of performance that is characterized by speech rate, pausing, hesitation, and repairs (Tavakoli et al., 2016; Lennon, 1990, 2000). As I discussed in Chapters I and II, utterance fluency, which measures "speech-planning difficulties that surface in utterances" (de Jong, Groenhout, Schoonen, & Hulstijn, 2015, p. 225), is the most researched type of fluency. De Jong, Steinel, Florijn, Schoonen, and Hulstjin (2013) and Segalowitz (2010) suggest that utterance fluency can be divided into three subconstructs of speed (i.e. speech rate/articulation rate), breakdown (i.e. silence/pausing), and repair (i.e. false starts/self-repairs/repetitions) fluencies. In another fluency framework, Skehan (2014) distinguishes between disfluencies that affect speech flow (i.e. pausing/reformulations) and those that negatively influence the speed (i.e. speech rate). However, Skehan's (2014) framework does not distinguish between breakdown and repair fluency.

According to Gorsuch (2011), disfluencies like "slow speech rate, false starts, and particularly [unnecessary] pauses" could lead to ITAs' "failure" in performance (p. 1). Accordingly, in the present study, I measured all three subconstructs of utterance fluency suggested by Segalowitz (2010) in the participants' monologic oral production. For speed fluency, articulation rate was utilized (Mirdamadi & de Jong, 2015; Witton-Davies, 2014). Articulation rate was calculated by dividing the number of syllables by the total speaking (or phonation) time (de Jong, 2013). Speaking time equals the “total time excluding the silent pauses” (de Jong et al., 2015, p.230). Following Tavakoli‘s (2016) suggestion, I decided to avoid using hybrid measures of utterance fluency (e.g. speech rate, phonation time, and mean length of run) to prevent any overlap between the measures. For instance, speech rate, as Tavakoli et al. (2016, p. 455) argue, “combine[s] pausing and speed aspects” of utterance fluency (see also de Jong et al., 2013). Breakdown fluency, in my study, was measured through frequency of silent pauses (Révész et al., 2014; Skehan, 2009).[7] Many of the fluency researchers (e.g. Révész et al., 2014; de Jong, 2013; de Jong et al., 2015) have unanimously proposed that the cut-off point should be 250 milliseconds (.25 seconds) for silent pauses. In order to analyze repair fluency, I used number of false starts and number of reformulations or self-repairs (de Jong et al., 2015; Tavakoli et al., 2016). The preliminary analysis of the transcripts indicated that there were more occurrences of reformulations over false starts. Additionally, some CAF studies (e.g. Révész et al., 2014) show that false starts have a stronger ‘predictive’ value for communicative adequacy. Therefore, both false starts and reformulations were investigated.

To summarize, I used different CAF measures to weigh the linguistic performance of the ITA candidates who took part in this study (see Table 6). To analyze syntactic complexity, measures of length and subordination were utilized. Mean T-unit length measured length and proportion of clauses to T-units helped me analyze subordination. Percentage of error free T-units and mean number of unintelligible words per 100 words measured accuracy. Articulation

[7] Some studies (e.g. Révész et al., 2014) have used the number of silent pauses ‘per 100 words’ or ‘per first 60 seconds’ of the production. However, since the degree of cognitive demand varies throughout the task and this might affect pausing (see Wood, 2006), I decided to count the number of silent pauses through the whole ITA monologic performance that lasted roughly 5 minutes for each candidate.

rate, number of silent pauses, and number of false starts and reformulations respectively weighed speed fluency, breakdown fluency, and repair fluency.

Table 6. *Summary of CAF measures used in this study*

Construct	Measure
Syntactic Complexity (C)	Length: Mean length of T-unit (MLT) Subordination: Proportion of clauses to T-units (CL/T)
Accuracy(A)	Percentage of error free T-units (PEFT) Mean number of unintelligible words per 100 words (NUIW)
Fluency (F)	Speed fluency: Articulation rate (AR) Breakdown fluency: Number of silent pauses (NSP) Repair fluency: Number of false starts (NF) Number of reformulations (NR)

3.4.4.4. Compensatory strategies

ITAs might employ certain compensatory strategies to compensate for the linguistic problems that affect their oral performance (Bailey, 1984; Elder, 2001). However, no studies in the literature on ITA education and assessment thoroughly investigate the type of compensatory strategies ITAs employ and whether any of these strategies contribute more to the communicative adequacy that fits the ITA proficiency framework. Therefore, it was difficult to decide on what strategies to look for in the videotaped performances. Moreover, as I argued in Chapter II, some of the strategies that might help normal L2 users communicate their intended meaning might not be compatible in ITA situations where linguistic proficiency and teaching content in L2 are integrated. For instance, in the ITA Test, ‘foreignizing’ or ‘code-switching to L1 or L3’ (Celce-Murcia, Dörnyei, & Thurrell, 1995) can be regarded as instances of incomprehensibility. Similarly, restructuring or repetition amidst speech could indicate lack of fluency. It is worthy to note that avoidance strategies were excluded due to the difficulty in finding relevant objective measures, as well as the fact that, in most of the recognized taxonomies, avoidance strategies belong to a separate category than compensatory strategies. The other two compensatory strategies that I initially intended to code in the data included approximation (i.e. using a similar, often more general, word to refer to an item, e.g. *smoke* for *steam*) and circumlocution (i.e. explaining a concept or item instead of using the exact lexical item as referent, e.g. *the thing that beeps when fire* for *fire alarm*). Preliminary analysis of the coded data indicated that the instances of approximation and circumlocution were almost none across the three groups; therefore, these measures were not included in the final analysis because they failed to capture the construct of compensatory strategies.

Considering the nature of ITA proficiency and relying on the classifications proposed by Celce-Murcia et al. (1995), Dörnyei and Scott (1997), and Faerch and Kasper (1983), I focused on studying the use of all-purpose words and nonlinguistic means (see Chapter II for definitions and examples). To be specific, frequency of occurrence for all-purpose words was counted in the presentation section of the test. As defined in the previous chapter, all-purpose words are generalized lexical items that are used to refer to a known concept or object when the exact word is either missing or cannot be retrieved by the interlocutor (e.g. *thing* or *stuff*).

To measure nonlinguistic means, an overall/holistic score of 0 to 5 was assigned to each performance. Nonlinguistic means that were evaluated in the presentations included visual aids (e.g. Microsoft PowerPoint™ or Prezi™ presentations), gestures, movements, and eye contact. In evaluating the use of visual aids, the focus was on the quality of the slides in terms of how busy they were or whether the font size, font family and the background color were appropriate for the context and readable to the audience (see Halleck & Moder, 1995). Additionally, proper eye contact referred to when the test taker would evenly look at the audience members without ignoring anyone in the room (see Hoekje, 2016). Concerning body language and gestures, the evaluation was tailored to the accepted social semiotics of North American classrooms at the academic level (see Pan, 2016). A context-based example would be using hands (or pointing devices) to communicate the content of a graph/table on the PowerPoint™ slide.

In conclusion, the same fellow researcher who helped with the coding process and I gave the scores based on an agreed-upon list of descriptors for each point. The descriptors can be seen in Table 7. The average of the two scores was the value that I used in the statistical analyses. The score on nonlinguistic means was validated performing a between-groups ANOVA that indicated a statistically significant effect at the alpha level of .05, $F(2, 18) = 4.45$, $p = .030$, $\eta 2 = .372$.

Table 7. *Descriptors of the score on nonlinguistic means*

Score	Description
4-5	Successful use of visual aids; evenly-established eye contact; appropriate gestures and body language
2-3	Acceptable use of visual aids but minor errors; established eye contact but some audience ignored; acceptable gestures and body language but some inappropriate stances
0-1	Serious errors in using visual aids; eye contact not established; Serious problems in gestures and body language

3.5. Procedures and data collection

To collect the oral data, the ITA Test performances were videotaped at the beginning of two consecutive semesters in 2015 and 2016 (see Appendix B for a sample screenshot of one of the performances)[8]. The data from the 2015 test were collected using a Canon™ Camcorder (Model Vixia HF R400) and the device used for the 2016 test was a GoPro™ (Model Hero 3+ Black Edition). The videotaped performances, then, were manually transcribed. One transcript from each category (i.e. Passed, Provisionally Passed, and Failed) was checked by another researcher[9], which yielded a total inter-transcriber agreement of 93%. Next, the transcribed data were coded relying on the chosen CAF and compensatory strategies measures. However, not the entire coding procedure was based on the transcripts. In particular, the use of nonlinguistic means was coded by watching the videos with muted sound. The sound of the videos were muted for coding this measure in order for the coders to focus exclusively on nonlinguistic, compensatory strategies. Each measure was coded twice. Inter-coder reliability was calculated for 43% of the data (3 participants in each category, 9 in total) using two-way intraclass correlation coefficient (see Feng, 2015). Table 8 shows the inter-coder reliability coefficients for the measures that were operationalized based on the transcripts.

Table 8. *Inter-coder reliability of the coded measures*

Coded measure	ICC
MLT	.92
CL/T	.87
PEFT	.92
NUIW	.88
NFS	.91
NR	89
APW	.92

ICC = Intraclass correlation coefficient; $p < .01$

[8] IRB protocol no. AS14137 (see Appendix D for the approval letter)

[9] The other researcher held a Ph.D. degree in Applied Linguistics and was familiar with the research on CAF/ITA.

Furthermore, I used a script by de Jong (2013) in Praat software (Boersma & Weenink, 2015) to analyze two of the fluency measures, namely articulation rate and number of silent pauses (see also de Jong & Wempe, 2009). I need to mention that I utilized Audacity software to filter the background noise in some of the extracted audio files for speed and breakdown fluency analysis in Praat. A sample test-taker profile from one of the Passed performances, along with the coding map and an excerpt from the coded transcript are presented in Appendix C.

3.6. Statistical analysis

In addition to inter-transcriber agreement (93%) and inter-coder reliability ($.88 < ICC < .92$, $p < .01$), inter-rater reliability was also calculated for the two ratings of the ITA candidate's oral performance that indicated a high correlation between the two sets of scores, $\kappa = .85$, $p < .01$. Then, the test of normality was run to ensure the data came from a normally distributed population.

In order to investigate the predictive power of CAF measures and compensatory strategies in ITA Test success, multiple regression analysis was conducted to indicate a model with the contributing variables with the highest predictive power. The first regression model used CAF measures as predictors across the ITA test scores, whereas the second model was based on measures of compensatory strategies as predictor variables. The obtained effect sizes determined the extent to which variation in ITA Test scores could be predicted by the variation in the independent variables. Moreover, one-way ANOVA and post hoc analysis (Tukey's HSD and Fisher's LSD) were used to pinpoint how contributing predictors differentiated between the three ITA groups. It is worth noting that test of collinearity was also run, which indicated that multicollinearity is not of concern.

To look for possible trade-off effects between CAF constructs, correlation analysis was used. To that end, the statistically significant correlations were reported.

Finally, the obtained regression coefficients were compared to find out which of the contributing predictors form the regression models have the strongest predictive power. The results of these statistical analyses will be explained in the next chapter.

CHAPTER IV.

RESULTS

4.1. Overview

This chapter summarizes the results of the statistical analyses that I ran in order to test the normality of the data, obtain descriptive statistics for each independent variable, and determine the predictive power of the variables through multiple regression analysis to answer the research questions of the study.

A number of independent (or predictor) variables were utilized in this study against the dependent variables of international teaching assistants' (ITA) proficiency groups including Passed (P), Provisionally Passed (PP), and Failed (F) candidates. To be specific, syntactic complexity was measured through mean length of T-unit (MLT) and proportion of clauses to T-units (Cl/T). Percentage of error free T-units (PEFT) and mean number of unintelligible words per 100 words (NUIW) were used to measure accuracy. To measure fluency, articulation rate (AR) for speed fluency, number of silent pauses (NSP) for breakdown fluency, and number of false starts (NF) and number of reformulations (NR) for repair fluency were used. All-purpose words (APW) and the score on nonlinguistic means (NLM) were used to analyze compensatory strategies (see Chapter III for the rationale behind selecting these measures).

4.2. Normality of the data

To test whether the data were obtained from a normally distributed sample, a test of normality was run in SPSS software (version 23) considering the nature of the data (i.e. continuous scaled data). The null hypothesis was that the distribution of the sample was not significantly different from a normal distribution. The Shapiro-Wilke W test indicated that, for all of the variables, the null hypothesis was accepted, and therefore, the distribution was normal and the sample fitted the assumption of normality (see Table 9). It is worth mentioning that Shapiro-Wilk is considered "an effective measure of normality" for smaller samples (Shapiro & Wilk, 1965, p. 602).

Table 9. *Results of the test of normality*

	Shapiro-Wilk		
Measure	Statistic	df	Sig.
MLT	.902	21	.062
CL/T	.959	21	.587
PEFT	.956	21	.532
NUIW	.902	21	.062
AR	.916	21	.108
NSP	.971	21	.820
NFS	.930	21	.196
NR	.916	21	.108
APW	.862	14[a]	.052
NLM	.909	21	.084

Note: $N = 21$; $p < .05$

[a] The frequency of APW was zero in the PP group.

4.3. Descriptive statistics

Across the three ITA groups, Table 10 summarizes the descriptive statistics on the predictor variables of the study including eight measures of CAF and two measures of compensatory strategies. I should emphasize that, after calculating the descriptive statistics, the values for the measures were first inspected to control for the outliers. In other words, values with more than three standard deviations (3 *SD*) from the mean were excluded from the analysis. Therefore, from the initial pool of 24 coded video-taped performances, 21 were utilized in the final inferential analyses.

4.4. RQ1: Predictive power of CAF constructs

The first research question of the study posed whether constructs of complexity, accuracy, and fluency (CAF) and their related measures predicted ITA candidates' successful performance in the ITA test, and if so, to what extent CAF constructs differentiate between ITA candidates based on their test performance. To answer this question multiple regression analysis was conducted, using SPSS (version 23). To be specific, multiple regression analysis was performed to determine a model indicating the CAF measures with the strongest predictive power through the obtained adjusted R^2 value. In case of statistical significance, one-way ANOVA and post-hoc analysis (i.e. Tukey's HSD and Fisher's LSD tests) were conducted to determine where the difference existed between the three ITA candidates' groups based on the ITA Test results. An alpha level of $p < .05$ was set for all the tests in the remainder of this chapter.

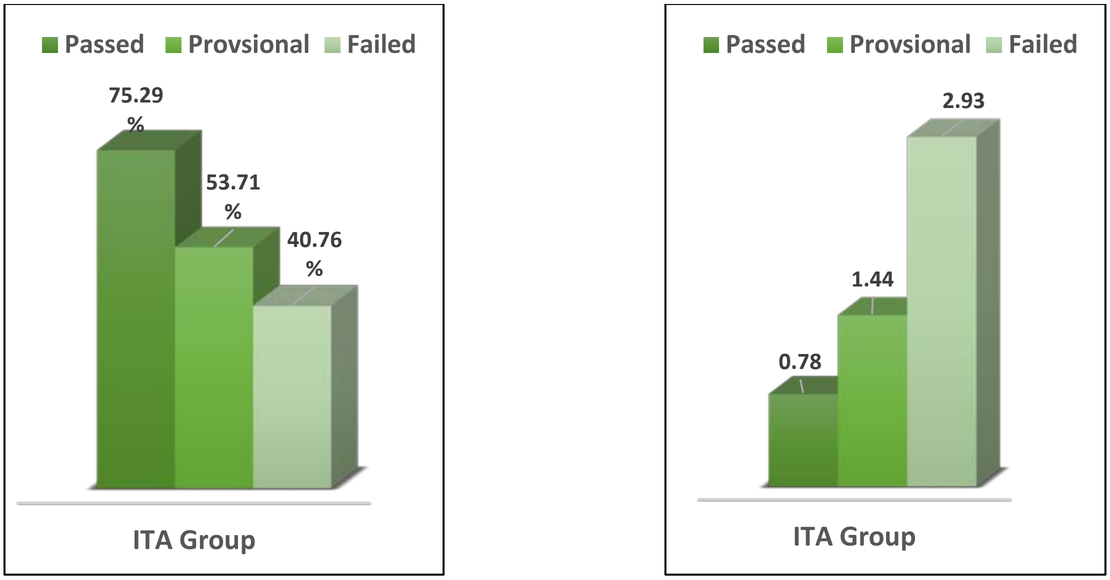
Table 10. *Descriptive statistics of the measures across the three ITA groups*

Measure	ITA Group	N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
MLT	Passed	7	10.05	1.42	.57	8.26	11.81
	Provisional	7	12.98	4.98	2.03	7.44	19.81
	Failed	7	11.71	2.09	.85	9.50	14.77
	Total	21	11.58	3.27	.77	7.44	19.81
Cl/T	Passed	7	1.35	.14	.05	1.17	1.61
	Provisional	7	1.32	.29	.12	1.00	1.83
	Failed	7	1.29	.14	.05	1.08	1.48
	Total	21	1.32	.19	.04	1.00	1.83
PEFT	Passed	7	75.29	8.32	3.39	62.50	86.96
	Provisional	7	53.71	14.80	6.04	33.33	66.66
	Failed	7	40.76	10.24	4.18	26.08	55.55
	Total	21	56.59	18.17	4.28	26.08	86.96
NUIW	Passed	7	.78	.54	.22	.00	1.66
	Provisional	7	1.44	.62	.25	.60	2.30
	Failed	7	2.93	2.01	.82	.60	6.30
	Total	21	1.72	1.49	.35	.00	6.30
AR	Passed	7	4.60	.74	.30	3.21	5.38
	Provisional	7	4.75	.36	.14	4.24	5.25
	Failed	7	4.43	.72	.29	3.35	5.38
	Total	21	4.59	.60	.14	3.21	5.38
NSP	Passed	7	68.33	23.64	9.65	41.00	101.00
	Provisional	7	94.83	27.11	11.06	64.00	143.00
	Failed	7	88.33	17.31	7.06	73.00	114.00
	Total	21	83.83	24.56	5.79	41.00	143.00
NFS	Passed	7	3.83	1.94	.79	1.00	6.00
	Provisional	7	2.33	1.36	.55	.00	4.00
	Failed	7	3.00	2.89	1.18	.00	7.00
	Total	21	3.05	2.12	.50	.00	7.00

NR	Passed	7	2.33	1.75	.71	1.00	5.00
	Provisional	7	7.66	3.98	1.62	3.00	13.00
	Failed	7	8.66	4.32	1.76	4.00	15.00
	Total	21	6.22	4.38	1.03	1.00	15.00
APW	Passed	7	2.00	2.44	1.00	.00	5.00
	Provisional	7	.00	.00	.00	.00	.00
	Failed	7	2.66	2.65	1.08	.00	7.00
	Total	21	1.55	2.28	.53	.00	7.00
NLM	Passed	7	4.12	.44	.17	3.50	4.75
	Provisional	7	3.91	.58	.23	3.00	4.50
	Failed	7	3.50	.44	.18	3.00	4.00
	Total	21	3.84	.53	.12	3.00	4.75

4.4.1. Result of the multiple regression analysis

A multiple regression analysis was performed using ITA candidates' test scores (250 to 300 for P, 240 to 249 for PP, and 239 or less for F) as the criterion and CAF measures as predictor variables conductive to determine whether ITA Test performance could be predicted as a function of any of the CAF measures used in this study. The analysis was found to be statistically significant (R^2 =.75, F(2,18) = 22.82, p < .000). Moreover, the results of the regression indicated that two predictors of PEFT and NUIW contributed to the model, and consequently, accounted for 72% of the variance in the ITA Test results, as indexed by the adjusted R^2 statistic. Figures 6 shows the means for PEFT (a) and NUIW (b) across the three test result groups of P, PP, and F. Additionally, Figure 7 and Figure 8 display the scatterplot of ITA Test scores for PEFT and NUIW respectively.



(a) (b)

Figure 6. The means for PEFT (a) and NUIW (b) across the three ITA test groups

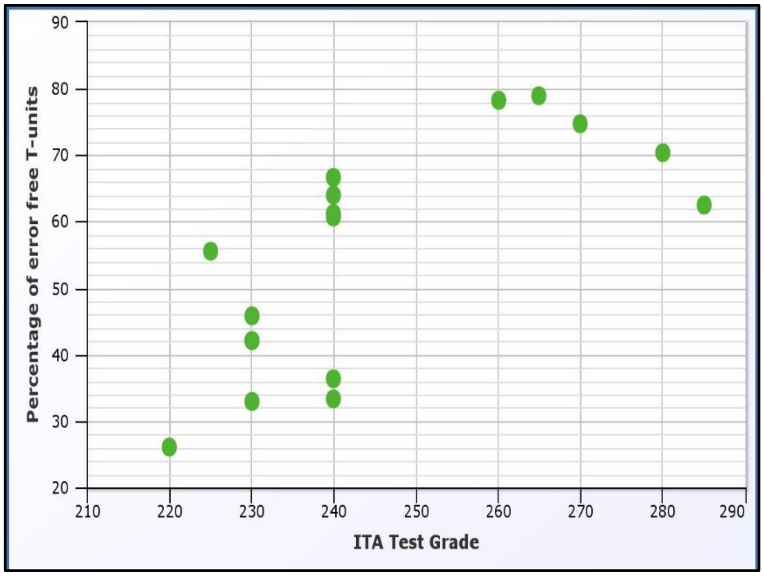


Figure 7. The scatterplot of ITA Test scores for PEFT

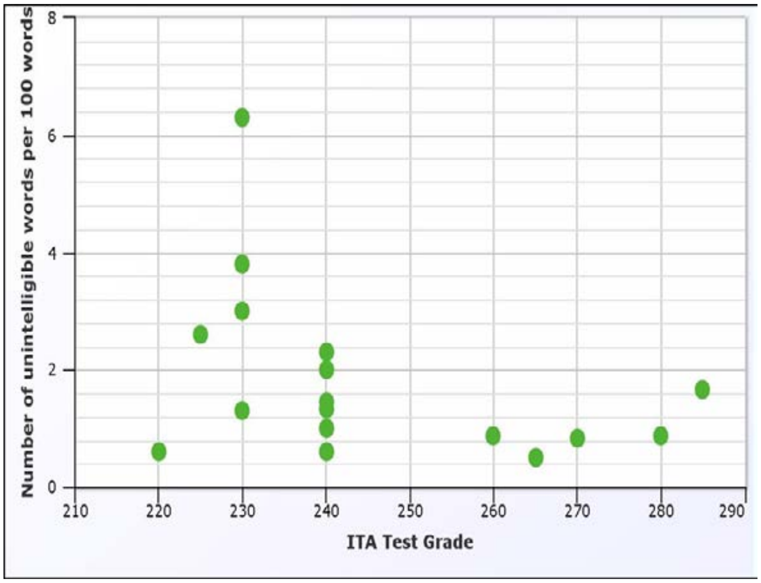


Figure 8. The scatterplot of ITA Test scores for NUIW

Indexed by the β value, PEFT (β = .669) was numerically a stronger independent predictor comparing to NUIW (β = -.364) when the unstandardized coefficients were compared. Moreover, the β value for PEFT was positive, that is, the increase in PEFT may subsidize an increase in the ITA Test score. On the contrary, the β value for NUIW was negative, meaning that an increase in NUIW could predict a decrease in the ITA Test score. However, in order to test this hypothesis, standardized β coefficients were compared in their real and in their absolute numerical values.[10] Consequently, PEFT was a stronger predictor of ITA Test result. To test further this hypothesis that PEFT is stronger than NUIW, all the values for the criterion (or the dependent variable) and the predictors (or independent variables) were transformed into Z-variates (i.e. standardized variables; see Gujarati, 2015). The multiple regression was run one more time with the Z-variates, which yielded the same results and the same β coefficients. In order to test the hypothesis that the standardized β weights of PEFT and NUIW were

[10] Absolute values were used here, since, due to the nature of some of the predictor variables, such as NUIW, the β coefficient could be a negative value.

significantly different, the 95% interval was estimated for each variable (Cumming, 2009). Since there was no overlap between the two β weights, it can be concluded that the larger β value (i.e. PEFT) had more predictive power than the smaller β value (i.e. NUIW) based on the results of this study.

4.4.2. Result of the post-hoc analysis

A one-way between groups ANOVA was used to compare the measure of PEFT across the three groups of P, PP, and F. The impact was found to be statistically significant, *F*(2, 18) = 13.92, *p* < .000. Post-hoc analysis using Tukey HSD and Fisher's LSD helped identify a significant difference between the P (*M* = 75.29, *SD* = 8.32) and the PP (*M* = 53.71, *SD* = 14.80) groups, as well as between the P (*M* = 75.29, *SD* = 8.32) and the F (*M* = 40.76, *SD* = 10.24) groups. Nonetheless, no significant difference was identified among the PP and the F groups (see Figure 8).

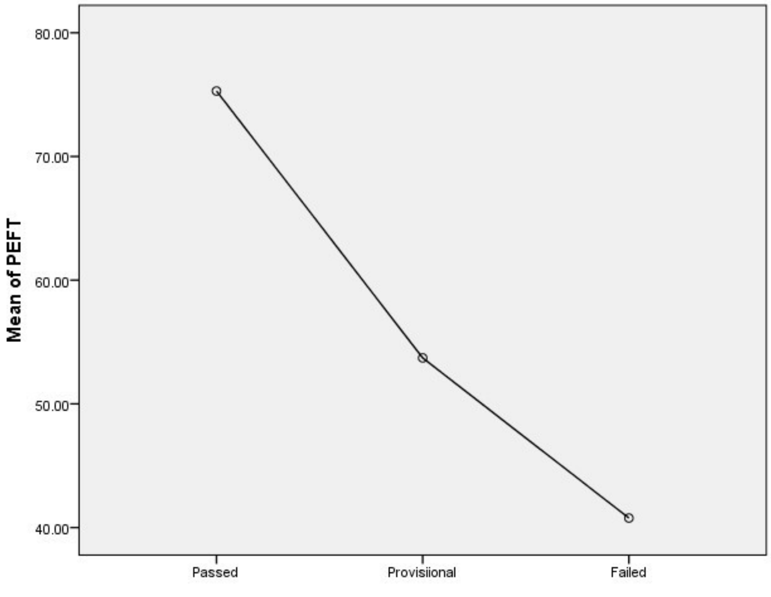


Figure 8. The means plot for PEFT

Similarly, a one-way ANOVA was conducted for NUIW across the three ITA groups. Consequently, a significant effect was found for NUIW, *F*(2, 18) = 4.57, *p* = .028. As displayed in Figure 10, post-hoc tests (i.e. Tukey HSD and Fisher's LSD) indicated that the P group (*M* = .78, *SD* = .54) had a significantly lower NUIW than the F group (*M* = 2.93, *SD* = 2.01). However, the mean differences among the P and PP groups, and the PP and F groups were not statistically significant.

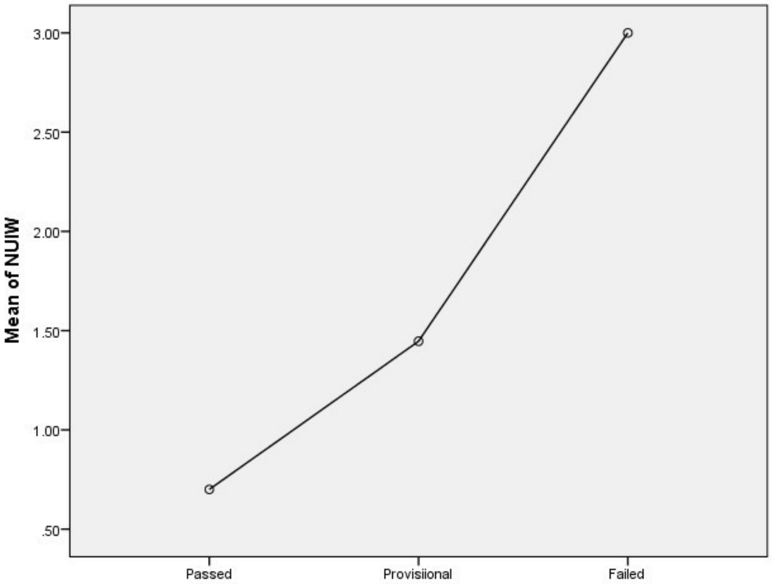


Figure 10. The means plot for NUIW

4.5. RQ2: The relationship between CAF constructs

The second research question of the study dealt with the relationship between CAF constructs. As shown in Table 11, all the statistically significant relationships between the constructs were negatively correlated, and the same time, relatively strong. To be specific, when measured by CL/T, complexity was negatively correlated with accuracy measured by NUIW (*r* = -.563, *p* < .05). In other words, increase in NUIW would negatively affect complexity. Furthermore, accuracy measured by PEFT had a negative correlation with fluency when measured by NR (*r* = -.625) with *p* < .01. That is, increase in NR as a measure of fluency would result in reduction of PEFT as a measure of accuracy. The obtained correlations clearly show that there is a competition among CAF constructs. That is to say, negative correlations suggest that there were trade-off effects between CAF constructs, namely between accuracy and syntactic complexity, and between accuracy and breakdown fluency.

Table 11. *Significant correlations between CAF measures (Layout by Vercellotti, 2015, p. 14)*

	Accuracy (NUIW)	Fluency (NR)	Fluency (NFS)
Complexity (Cl/T)	-.563*		
Accuracy (PEFT)		-.625**	
Fluency (AR)			-.562*

* Significant at *p* < .05
** Significant at *p* < .01

Another interesting correlation that was found was within the construct of fluency. The negative correlation between two fluency measures of AR and NFS was significant (*r* = -.562, *p* < .05), meaning that escalation of NFS would lead to decrease in AR. The rest of the calculated correlations were either modest or weak and statistically not significant. Thus, it can be concluded that there is no overlap among the rest of the CAF measures in this study, and consequently, such measures assess different aspects of the ITA candidates' linguistic performance in the ITA Test.

4.6. RQ3: Predictive power of compensatory strategies

The third research question of the study concerned the power of compensatory strategies in predicting the performance of candidates in the ITA Test. Moreover, it inquires the extent to which compensatory strategies differentiate between ITA candidates based on their test performance. To that end, multiple regression analysis was conducted to produce a model identifying the contributing independent variables to ITA Test performance. Then, the analysis was followed by Tukey's HSD and Fisher's LSD post-hoc tests to indicated where the difference existed between the three ITA groups.

4.6.1. Result of the multiple regression analysis

Similar to the analysis used in RQ1 (see section 4.4), a multiple regression analysis was performed using ITA candidates' test scores as the criterion and compensatory strategy measures (i.e. APW and NLM) as predictor variables. Multiple regression was used in order to find out whether ITA Test performance could be predicted as a function of any of the compensatory strategies used in this study. The multiple regression model produced R^2 = .24, F(1,19) = 5.04, p = .039. Based on the model, NLM was found to be the contributing variable that, as indexed by the adjusted R^2 statistic, accounted for about 19% of the variance in the ITA Test results (see Figure 11). As discussed in the last chapter, NLM was measured by assigning a score between 0 and 5 to each test taker by the two coders.

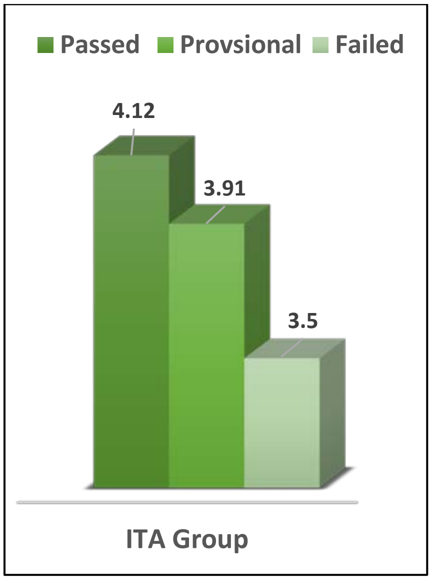


Figure 11. The means for NLM across the three ITA test groups

4.6.2. Result of the post-hoc analysis

Post-hoc analysis using Tukey HSD and Fisher’s LSD helped identify possible significant differences between the three groups of P, PP, and F in terms of NLM. Although Tukey HSD test did not signify any differences between the groups, Fisher’s LSD identified a statistically significant difference between the P (*M* = 4.12, *SD* = .44) and the F (*M* = 3.50, *SD* = .44) groups. However, no significant difference was found between the P and the PP groups, as well as between the PP and the F groups (see Figure 12).

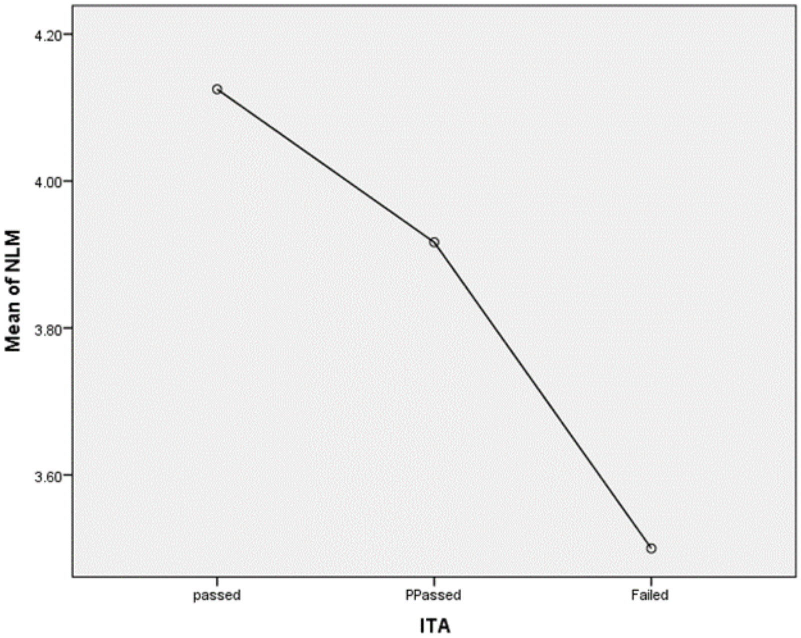


Figure 12. The means plot for NLM

Here, I need to make a case for using and relying on Fisher’s LSD as a useful and credible post-hoc test. It has been both analytically (e.g. Hayter, 1986) and empirically (e.g. Seaman, Levin, Serlin, 1991) shown that LSD works well when three means (or groups) are compared, such as the three ITA groups in the present study (see also Levin, Serlin, & Seaman, 1994). According to Seaman et al. (1991), comparing to Tukey's HSD, Fisher's LSD is about 8% more powerful as a post-hoc test.

4.7. RQ4: Comparing CAF constructs and compensatory strategies

The last research question of the study probed for the inclusive comparison of CAF measures and compensatory strategies, which could be answered form two perspectives: a) comparing the obtained R^2 statistics of the produced models for CAF and compensatory strategy measures; b) comparing the obtained β values of the modeled (or contributing) predictor variables. However, before comparing the β values, multicollinearity diagnostics was run to see whether the predictor variables (i.e. all the independent variables used in the multiple regression analyses) met the assumption of collinearity. With all the Variance Inflation Factor (VIF) values less than 10 and the Tolerance greater than .1, the tests indicated that multicollinearity is not a concern. In other words, correlations between the independent variables did not adversely affect the regression estimates in the conducted analyses.

4.7.1. Comparing the obtained R^2 statistics

As discussed in the previous sections, the obtained R^2 statistic for each of the produced models in the two multiple regression analyses that were run for CAF constructs and compensatory strategies indicated the variances that were found in the ITA Test results accounted for by the contributing predictors. On the one hand, the produced model through the first regression analysis showed that PEFT and NUIW are the contributing predictors to the model when CAF measures are predictor variables. More importantly, based on the adjusted R^2, these two accuracy measures predicted 72% of the variance in the ITA candidates' performance. On the other hand, the second regression analysis was conducted with compensatory strategies as the predictors. The results showed that NLM was the contributing predictor. Furthermore, the adjusted R^2 indexed that 19% of the variance in the ITA candidates' performance could be predicted by NLM. Therefore, it can be concluded that by comparing the CAF measures of PEFT and NUIW predicted a huge percentage of the variance, that is, 72%, whereas the compensatory strategy of NLM accounted for only 19% of the variance.

4.7.2. Comparing the obtained β values

In terms of comparing β values, an analysis similar to the one used in the second half of section 4.4.1 was sued. To be precise, the standardized β weights for the modeled predictors, which were resulted from the two conducted regression analyses based on Z-variates, were compared. In order to test the hypothesis that the standardized β weights of PEFT, NUIW, and NLM were significantly different, the 95% interval was estimated for each variable (Cumming, 2009). Prior to testing the hypothesis and investigating the overlap between the values, the numerical value of the standardized β weights were in this order: β_{PEFT} (= .669) > β_{NLM} (= .489) > β_{NUIW} (= -.364). As a result, no overlap was found between the three β weights. Thus, PEFT was the strongest predictor, followed by NLM, and finally, NUIW.

4.8. Summary of the results[11]

Overall, the results of the study revealed that both accuracy measures used in this study, that is, percentage of error-free T-units and number of unintelligible words per 100 words, outperformed their complexity and fluency counterparts in predicting the test takers' ITA Test performance. In addition, regarding compensatory strategies, nonlinguistic means was found to

[11] To increase the readability of the last section of this chapter, the complete name of the measures were used instead of their related abbreviations.

be another strong predictor of ITA Test performance when compared to all-purpose words. The correlations between the eight CAF measures of this study indicated that there is a trade-off between complexity (proportion of clauses to T-units) and accuracy (number of unintelligible words per 100 words). Another trade-off was found between accuracy (percentage of error-free T-units) and fluency (number of reformulations). Last but not least, the accuracy measures, altogether, accounted for a larger variance in the performances across the three ITA groups, whereas, when β weights were compared, percentage of error-free T-units was the strongest predictor. The second stronger predictor was nonlinguistic means, followed by number of unintelligible words per 100 words.

In the next chapter, I will present an in-depth discussion of the results based on the literature on ITA assessment, the CAF framework, and compensatory strategies. Additionally, I will discuss the assessment and pedagogical implications pertaining to the results of my study. Finally, I will propose the potential topics as future directions for the ITA assessment research based on CAF constructs and ITA-related strategies.

CHAPTER V.

DISCUSSION

5.1. Overview

In light of the research objectives, Chapter V elaborates on the findings of the study, which were derived from the results of the statistical analysis. Additionally, the findings will be discussed with respect to the literature on complexity, accuracy, and fluency (CAF) and international teaching assistant (ITA) proficiency. To conclude, implications for the stakeholders in ITA education and assessment, as well as limitations of the study will be introduced. To review, I formulated the following four research questions to address the objectives of the study:

RQ1: Do CAF constructs and their related measures predict ITA candidates' success in an ITA test? If so, what is the extent to which CAF constructs differentiate between ITA candidates based on their test performance?

RQ2: Is there a relationship between CAF constructs? If yes, how do CAF constructs influence each other when the task condition is not a variable?

RQ3: Do compensatory strategies predict ITA candidates' success in an ITA test? If so, to what extent do compensatory strategies differentiate between ITA candidates based on their test performance?

RQ4: Upon comparing CAF measures and compensatory strategies, which one is a better predictor of ITA candidates' success in the ITA test?

5.2. Discussion of the answers to the research questions

5.2.1. Predictive power of CAF constructs

The first section of RQ1 was concerned with whether CAF could predict test takers' performance in the ITA Test. The results of the statistical analysis confirmed that CAF measures can predict the candidates' ITA proficiency if measured by the ITA test. However, not all the CAF measures were strong predictors. To be specific, only accuracy significantly predicted the ITA Test result, whereas complexity and fluency did not have a strong predictive power. As I reviewed in Chapter II, according to Foster and Wigglesworth (2016), accuracy is a reliable indicator of communicative competence. The first finding of the study verifies this very characteristic of accuracy. Additionally, a possible explanation for the finding that accuracy was the only construct that predicted the ITAs' performance might be because it is, among CAF constructs, the "most internally coherent construct" (Pallotti, 2009, p. 529). This finding is also in agreement with Halleck (1995), who reported that accuracy impacted proficiency more than complexity in her subjects' performance in the Oral Proficiency Interview (OPI). However, Halleck's study did not involve fluency, and it was in a Chinese EFL context. Moreover, this finding adds to the confusion about the effect that pre-planning (although not the focus of this study) has on CAF in L2 oral production. Yuan and Ellis (2003), for example, found that pre-planning enhanced syntactic complexity. Assuming that the test-takers in the ITA Test had unlimited pre-planning time, their linguistic production was significantly influenced by accuracy, rather than complexity and fluency. In another study on the predictive power of CAF measures, Révész, Ekiert, and Torgersen (2014) concluded that fluency was the strongest predictor of communicative adequacy of advanced learners. This is in contrast with the first finding of the study that suggests accuracy was the strongest predictor.

Moreover, the variance in the test results was mostly accounted for by the two global accuracy measures used in the study, that is, percentage of error-free T-units (in short, error-free T-units) and number of unintelligible words per 100 words (in short, number of unintelligible words). Error-free units of segmentation are recurrently used in the research on accuracy and CAF as reliable measures (see Foster, Tonkyn, & Wigglesworth, 2000). Number of unintelligible words, though, is a measure that was validated and used for the first time in this study. As I explained in the Methodology chapter, during the coding process, careful analysis of the transcriptions highlighted frequent occurrence of words that were unintelligible to the transcribers in the Provisionally Passed and Failed performances. Since Tonkyn (2012) suggests that accuracy and comprehensibility are interwoven concepts, number of unintelligible words was categorized as an accuracy measure. Surprisingly, this invented measure turned out to be a strong predictor of the ITA candidates' oral performance. Thus, it is logical to assume that number of unintelligible words can serve as a valid and reliable accuracy measure when global errors, both syntactic and lexical, are of concern (Foster & Wigglesworth, 2016). Nevertheless, attention should be paid to the fact that this measure is applicable to a research design that capitalizes on transcribed data for analysis.

In terms of differentiating between the ITA groups, number of unintelligible words was higher in the Failed group compared to the other two groups. The Passed ITAs had the least unintelligible speech among the three groups. Nonetheless, the distance between the Passed and the Failed group was the only statistically significant difference, indicating a meaningful proficiency level change between such groups. Regarding error-free T-units, Passed ITAs performed significantly more accurately (or error-free) than the Provisionally Passed and the Failed candidates. However, although the Provisionally Passed ITAs' performance was more error-free than their Failed counterparts, this difference was not significant. Thus, one can conclude that for both accuracy measures, the Provisionally Passed and the Failed ITAs do not differ significantly in their linguistic performance. Research on other accuracy measures, though, yields a rather different conclusion. For instance, Inoue (2016, p. 497) introduced the number of errors per 100 words as the “most valid measure of accuracy” when the proficiency levels include mostly beginners and intermediates. One important finding of Inoue's study is the role that denominators (i.e. 100 words, units, etc.) play in the distribution of errors in the oral performances. In this study, 100 words and T-units served as the denominators in measuring accuracy.

5.2.2. Competition between CAF constructs

The most theoretically motivated research objective of this study was to investigate the correlation/competition between CAF constructs and their measures when the task is oral performance in an ITA proficiency assessment. It is worth mentioning that the existing relationships between the measures were studied irrespective of their predictive weight with regard to the final outcome of the ITA Test performance. I argued in the previous chapters that there are somewhat opposing views of how cognitive demand influences the relationship between CAF constructs. Skehan's (1998) Limited Capacity Hypothesis asserts that, due to limitations in attentional capacity and working memory, L2 performance is affected by trade-off effects between CAF components. Robinson's (2003, 2005) Cognition Hypothesis, on the other hand, argues for the importance of task characteristics in shaping the relationship between CAF constructs. More precisely, Robinson (2011) claims that easy monologic tasks promote fluency in L2 oral production, whereas difficult monologic tasks facilitate complexity and accuracy. Furthermore, Robinson (2011) believes that syntactic complexity is reduced in dialogic and interactive tasks. Finally, according to de Bot's (2008) dynamic systems theory and Larsen-Freeman and Cameron's (2008) complexity theory, there is no causal relationship between CAF components, and instead, the components are all interconnected (see also Vercellotti, 2015).

The second research question, essentially, put these hypotheses to test in the domain-specific performance of the ITA candidates. Based on the results, meaningful trade-off effects were found between certain measures. To be specific, syntactic complexity and accuracy negatively correlated when measured by proportion of clauses to T-units and number of unintelligible words respectively. In other words, the ITAs who tried to produce more complex language by using more subordination did so by sacrificing accurate language, and therefore, by producing unintelligible utterances. This finding is in agreement with Skehan's (1998) idea of a limited cognitive capacity, indicating a trade-off between syntactic complexity in the form of subordination and accuracy in the form of unintelligibility. This is, nonetheless, in contrast with Robinson's (2011) claim that difficult monologic tasks promote complexity and accuracy but not fluency. Given the assumption that the monologic section of the ITA Test can be categorized as a complex task, complexity and accuracy did not get promoted alongside in this case. Moreover, it was found that more fluent language came at the price of accuracy for the ITA candidates in this study. To be precise, the candidates sacrificed error-free language to speak with fewer reformulations. This finding, once again, verifies the trade-off effect proposed by the Limited Capacity Hypothesis. The last noteworthy competition was within the construct of fluency. The ITAs in this study compensated for speed by pledging less repair in their oral production. That is, the articulation rate significantly increased as the frequency of false starts decreased. This negative correlation, however, is already recognized in the literature on L2 fluency (e.g. Lennon, 2000). Previously, Malicka (2014) found an internal trade-off within the construct of syntactic complexity in monologic oral production when the measures included number of words per AS-unit[12] and number of words per clause. In this study, though, the trade-off effect between syntactic complexity measures was negligible. The findings of this study related to the competition between CAF constructs is also in contrast with Vercellotti's (2015) findings. Specifically, Vercellotti found that all CAF measures in her study positively interacted with each other, experiencing an all-inclusive growth over time. It is needless to say that Vercellotti's findings were obtained longitudinally within a developmental framework, whereas this study was focused on a one-time test performance.

To summarize, when the task was oral performance in the ITA Test, the competition between CAF constructs partially confirmed Skehan's (1998) Limited Capacity Hypothesis, but partly contradicted Robinson's (2003, 2011) Cognition Hypothesis. With regard to Dynamic systems theory (de Bot, 2008) and complexity theory (Larsen-Freeman & Cameron, 2008), the findings did not suggest a systematic 'interrelatedness', and on the contrary, some causal and trade-off effects were found.

5.2.3. Predictive power of compensatory strategies

Along with Bailey (1984), Elder (2001), and Halleck and Moder (1995), this study argued that the construct of ITA proficiency entails more than merely the linguistic components, that is, the subconstructs measurable through the CAF triad. I stressed in Chapter III that although there are studies testifying to the significance of strategic competence and

[12] Refer to Chapter I for the definition of AS-units.

compensatory strategies for ITAs (e.g. Ard, 1987; Halleck & Moder, 1995), there is a lack of research evidence on exactly what strategies help ITAs compensate for their linguistic deficiencies. Thus, having the contextual nature of the ITA proficiency in mind, the compensatory strategies that I coded in the ITA performances were based on the research within the general communicative competence framework (e.g. Celce-Murcia, Dörnyei, & Thurrell, 1995; Dörnyei & Scott, 1997; Natakatani, 2006).

To investigate whether compensatory strategies predicted the candidates' performance in the ITA Test, certain strategies were chosen to be analyzed. Specifically, the holistic score on nonlinguistic means (in short, nonlinguistic means) and frequency of all-purpose words (in short, all-purpose words) were used in the final analysis of the performances during the monologic task. The results of the statistical analysis confirmed that, similar to CAF measures, compensatory strategies also contributed to the candidates' performance. This first finding is in line with previous arguments with regard to the role of compensatory strategies in ITA proficiency (e.g. Ard, 1987; Bailey, 1984; Elder, 1993, 2001; Halleck & Moder, 1995). More precisely, it was the nonlinguistic means that predicted the ITA Test result, whereas all-purpose words did not have a strong predictive power. All-purpose words did not occur frequently in the presentations or the Q&A section (see Table 8 in Chapter IV), indicating that the ITAs did not use this strategy against a potential lack of cognitive access to their mental lexicon. Another justification for the fact that the ITAs did not employ all-purpose words to a great extent might be the impact that planning time and task type had on their performance at the lexical level. Formerly, Yuan and Ellis (2003) reported on the positive effect of pre-planning on lexical variety in L2 monologic oral speech (see also section 5.2.1 in this chapter).

In terms of the differences between the three ITA groups, Passed ITAs used nonlinguistic means significantly more than their Failed counterparts. However, the difference in using nonlinguistic compensatory strategies was not considerable between Passed and Provisionally Passed and between Provisionally Passed and Failed ITAs. In other words, predicting the performance in the ITA Test, use of compensatory strategies distinguished Passed ITAs from Failed candidates. This very finding is in agreement with Halleck and Moder (1995, p. 733) who argue that “Compensatory teaching strategies, which enable more proficient students to overcome linguistic weaknesses, do not have a strong effect for less proficient learners.” In the case of the ITAs in this study, utilizing nonlinguistic means such as incorporating visual aids in the presentation, establishing eye contact, and employment of domain-appropriate gestures predicted a passing score on the ITA Test. On the contrary, the candidates who failed the test did not use this group of compensatory strategies to the extent that could have contributed to a successful test performance.

Before I conclude this section, there are two side observations based on the data that may be worthy to note. Firstly, it is interesting that none of the ITA candidates used circumlocution and approximation at all during either the monologic or the dialogic task (although this task was not used in the analysis). Nonetheless, this compensatory strategy is often used by other L2 speakers with high levels of proficiency and in other communicative situations (Dörnyei & Scott, 1997). One possible explanation for the absence of circumlocution in the ITAs’ performances might be the technical content of the presentations that contains highly frequent, field-specific jargon. Secondly, none of the candidates used anything but Microsoft PowerPoint™ as the main visual aid to present their content (see Appendix B). Based on my experience as an ITA Test rater, I had rated presenters that used other tools such as the blackboard or Prezi™. This indicates the continuous popularity of PowerPoint™ presentations in academic teaching, a point that requires attention from ITA program directors and educators. It seems like that this prevalent method of presentation has found its way in academic in the States to the extent that the new, less branded methods are yet to be used extensively by presenters.

5.2.4. Comparing the predictive power of CAF constructs and compensatory strategies

RQ4 was concerned with the predictive power of the contributing measures to the ITA Test performance. I concluded in the previous chapter that the variance in the test results was mostly accounted for by the accuracy measures. To be specific, number of unintelligible words and error-free T-units predicted 72% of the variance in the ITA candidates' performance. The high percentage of variance explained by the accuracy measures used in this study revealed the significant role of the linguistic components, in this case accuracy, of a performance test such as the ITA Test. The remaining variance was mostly (i.e. 19% of the variance) predicted by a compensatory strategy measure, that is, nonlinguistic means. Therefore, altogether, error-free T-units, number of unintelligible words, and nonlinguistic means hold a huge predictive power with regard to the ITA Test performance. Further analysis of the results yielded a clearer understanding of the difference between the predictive power of each of these measures. Precisely, comparing the predictive power of the contributing measures showed that error-free T-units was the strongest predictor of test performance for the ITA candidates in this study. Furthermore, nonlinguistic means were the second strongest predictor, followed by number of unintelligible words as the third strongest measure with predictive power.

It can be inferred from the discussion above that in the debate over the construct validation of ITA proficiency, attention should be paid to the potential salient contribution of accuracy, if globally measured by number of unintelligible words and error-free T-units, and compensatory strategies, if holistically measured by nonlinguistic means. This finding does not undermine the role that other CAF and compensatory strategy measures, as well as other contributing factors might play in validating the construct of ITA proficiency. This issue will be discussed further in the section on the limitations of this study and future directions for researchers.

5.3. General discussion and testing/teaching implications

With the increase in the number of ITAs, their screening, and therefore, training has gained crucial significance (Choi, 2017). In this study, I argued that popular tests of oral proficiency that are used to admit students to graduate programs fall short to assess the particular L2 skills that are required to perform TA duties. In particular, strategic competence that ITAs need in order to compensate for their linguistic problems cannot be reliably measured through such tests (Elder, 1993, 2001; Halleck & Moder, 1995). Even for the non-TA situations, Brooks and Swain (2014, p. 353) cast doubt on the "validity argument of the Speaking section of the TOEFL iBT" when it comes to international graduate students. As a result, the operationalization of ITA proficiency as an independent construct might make it possible for the stakeholders to benefit from the results of a test instrument that thoroughly captures what is actually supposed to be measured in an ITA's language. Using an in-house performance test exclusively developed to screen ITA candidates, this research project mainly aimed at investigating the extent to which linguistic components (measured by CAF) and strategic competence (measured by compensatory strategies) contribute to predicting ITA proficiency. This goal was set in spite of the absence of research that implements CAF in the ITA proficiency framework. Coupled with the absence of research using CAF in ITA assessment, no studies, so far, has ever researched the type of compensatory strategies ITAs employ in their performance.

As the findings suggest, both linguistic components and strategic competence might predict ITA proficiency when the test is performance-based and domain-specific (see Bowden, 2016). This is enlightening for the decision makers, such as ITA program directors, concerning the development of ITA screening tests that tap both the linguistic competence and the type of strategic competence compatible with TAs' linguistic involvement. Furthermore, training of the raters should be done in harmony with the development of a valid ITA test. At the linguistic level, for instance, this study revealed that the raters were more sensitive towards accuracy and not complexity or fluency. Even what followed accuracy, in terms of the predictive power, was strategic competence measured through the use of compensatory strategies. Here, I would like to echo what Elder (2001) emphasizes about L2 speaker teacher proficiency, which resembles the ITA situation to a great extent:

> "The construct of teacher proficiency, as operationalized in these performance-based measures of teacher proficiency, is clearly multidimensional, and this poses problems for the interpretation and reporting of performance. One solution to this problem would be to separate the purely linguistic and the more classroom-specific aspects of performance in our reporting (Elder, 2001, p. 163)."

In line with Elder's argument, one can claim that ITA proficiency is a multifaceted construct, and thus, it needs to be approached with the same central premise in mind. Similarly, it is significant for administrators of tests of ITA proficiency to report the test results in a componential way to the training (or ESL) course instructors, so the classroom practices can be tailored to the ITAs' specific needs (Saif, 2002). In the case of the main instrument of this study, the ITA Test, the candidates who did not pass the test were supposed to enroll in remedial courses to be able to retake the test and get certified to teach. Specifically, the students who provisionally passed the test were supposed to take a remedial course on ITA strategies and skills, that is, “classroom-specific aspects of performance”, while the students who failed were supposed to enroll in a course on “the purely linguistic” issues of performance (Elder, 2001, p. 163). Perhaps a modification to the ITA Test that might make the interpretation of the results easier for the stakeholders would be to remove the labels for the two groups of Provisionally Passed and Failed. Instead, the raters’ decision based on the performance might entail the type of deficiency that that ITAs’ proficiency suffers from, that is, whether it lacks linguistic competence, the knowledge of compensatory strategies, or both. In the case of the provisionally passed ITAs, although they had to take the remedial course in ITA skills, they could still be recruited as TAs by their departments. Nonetheless, ITAs who failed not only were supposed to take the course in grammar and pronunciation, but also they could not be employed as a TA. Such a decision could be adjusted in a way that students who suffer from deficiencies in both language and strategies are not granted the TAship.

Regarding the teaching implication, the findings once again emphasized the salience of improving accuracy in L2 oral performance. Foster and Wigglesworth (2016, p. 98) stress the central role of “accuracy in performance” in the learner’s success. It is on ITA program directors and educators to implement teaching lexical and grammatical accuracy as well as stressing the importance of intelligibility in the ITA-related curricula. Also, I argue along with Halleck and Moder (1995) that training ITA candidates in compensatory strategies should be done with care as long as the reason that a candidate was not cleared to teach was primarily strategic incompetence. The same condition applies to the candidates whose incompetency mainly stems from linguistic shortcomings. In other words, candidates whose oral production is not complex, inaccurate (in the case of this study), or disfluent may not benefit from a training course that is centered largely on strategies rather than language issues (see also Jenkins & Parra, 2003). ITA educators ought to conduct a needs analysis on the type of strategies that their strategically incompetent learners require to master based on the teaching context, scope of responsibilities, and the type of educational technologies that are available at the institutional level.

5.4. Limitations of this study and future directions

For certain, this research project dealt with certain limitations in its design, data collection, and interpretation of the results. First and foremost, this study, out of many, used only a certain number of measures to evaluate CAF in the ITA candidates' oral performance. I employed these eight measures primarily based on the nature of the oral performance in which the candidates produced L2 speech in a particular context of teaching academic content. For instance, measures of lexical variety and lexical complexity were not used given the different topics the candidates had chosen for their ITA Test performance (see Bulté, 2007; McCarthy & Jarvis, 2010; Vercellotti, 2015). Future studies can implement other CAF measures in their design provided that the use of the measure is justified on ITA grounds.

Furthermore, I relied solely on general CAF measures since previous research shows that such measures, as opposed to specific measures of CAF, yield better results when the focus in not on the production of a specific form (see Housen & Kuiken, 2009). Undoubtedly, future research on CAF and ITA proficiency can study specific measures (or subconstructs) of CAF in the ITA context. To specifically measure syntactic complexity, for instance, one can use number of finite verb phrases per a production unit in ITAs' oral speech (see Bulté & Housen, 2012 for a review of complexity measures). Focusing on a particular verb tense as the base for a specific grammatical accuracy measure would be another example (see Foster & Wigglesworth, 2016 for a review of accuracy measures). Finally, a specific measure of fluency that might fit the ITA proficiency agenda can be number of syllables per a proportion of the speech time (see Witton-Davis, 2014 for a review of fluency measures).

Another limitation of the present study was to use only one unit of production, that is, the T-unit in calculating the complexity measures as well as one of the accuracy measures. As I explained in Chapter III, T-units were utilized since the main focus was on the monologic task of the ITA Test. However, if the focus were on the dialogic task, T-units might have failed to account for the often broken language in the Q&A. A future research project on the dialogic interaction of the ITAs can use As-units, since such data “contain many nonsyntactic segments (Norris & Ortega, 2009, p. 560).”

There were also limitations caused by the main instrument of data collection in this study. The ITA Test had a holistic rubric, which made it difficult for reliable correlations to be made between the existing general subsets of the rubric and the contributing variables recognized by the regression models. In addition, I was restricted in terms of controlling for the possible variations between the raters. The performances were videotaped in two different semesters across different rooms. Each room had two raters who might have had different degrees of sensitivity to each of the CAF constructs and compensatory strategies. However, as I discussed in Chapter III, the raters had similar, vigorous training by the same person who was the director of the ITA program, and the high inter-rater reliability between the scores was a testimony to this matter. Another variable that the nature of the ITA Test did not allow to control for was the topic variation (see Papajohn, 1999). The analyzed presentations had dissimilar topics, although they were all related to a subject teachable at the undergraduate level at a North American college or university. Nonetheless, there is always a possibility that the raters enjoy or relate to a topic more than the other due to unobservable reasons. One possible, but hard-to-be-practical, solution to this potential problem is for future researchers to narrow down the participants to those who come from the same graduate program.

Finally yet importantly, this study did not distinguish between demographic characteristics of the participants. In other words, despite the fact that the ITA candidates were all graduate students and applicants for TA positions, due to limitations in the number of ITAs who were taking the test for the first time, this study did not differentiate them based on their age, sex, language background, and the degree they were seeking (i.e. Master's or Ph.D.). Other studies on the operationalization of ITA proficiency might draw a clearer picture of this construct by focusing on possible, meaningful differences in age, sex, L1, and the degree category.

5.5. Conclusion

The focus of this study was to fill the gap in the existing literature on operationalization of the construct of ITA proficiency based on the premise that ITAs' oral performance needs to be assessed independent from general L2 proficiency (Elder, 2001; Gorsuch, 2011; Halleck & Moder, 1995). To that end, CAF measures, as reliable descriptors of L2 performance (Housen, Kuiken, & Vedder, 2012), and compensatory strategies, as indicators of strategic competence (Dörnyei & Scott, 1997) were compared to see whether they predicted ITAs' performance in a domain-specific test, namely, the ITA Test. Performances of 21 ITA candidates who took the ITA Test were videotaped and analyzed using eight measures of CAF as well as two measures of compensatory strategies. To be specific, to measure syntactic complexity, mean length of T-unit and proportion of clauses to T-units were used. Accuracy, on the other hand, was measured by percentage of error free T-units and mean number of unintelligible words per 100 words. The latter accuracy measure was validated and used for the first time in this study to account for unintelligibility in the ITAs' produced speech. To measure fluency, articulation rate for speed fluency, number of silent pauses for breakdown fluency, and number of false starts and number of reformulations for repair fluency were employed. Frequency of all-purpose words and the score on nonlinguistic means were used to quantify the use of compensatory strategies. The results of the multiple regression analyses indicated that, altogether, both accuracy measures (i.e. percentage of error free T-units and mean number of unintelligible words) and one of the compensatory strategy measures (i.e. the score on nonlinguistic means) accounted for 91% of the variance in the test results. Comparing the predictive power of these measures revealed that the strongest predictor was percentage of error free T-units. Nonlinguistic means and unintelligible words were, respectively, the second and third strongest predictive measures. Finally, analysis of the correlations between the CAF measures showed that there is a trade-off effect between some of the measures confirming Skehan's (1998) Limited Capacity Hypothesis. Overall, the findings suggest that the final decision about provisionally passed and failed test takers needs to be made with care depending on the nature of the incompetence that causes the candidate not to pass the test. Also, ITA educators and program directors need to address accuracy and paralinguistic strategies in syllabus design and course planning. Without a doubt, the discussion about construct validation of ITA proficiency is far from over, and more research evidence is required to support the arguments in favor of the validity of in-house, domain-specific tests such as the ITA Test. This study, however, shed light on some of the potential contributors to the particular construct of ITA proficiency in its monologic aspect.

REFERENCES

Ard, J. (1987). The foreign TA problem from an acquisition-theoretic point of view. *English for Specific Purposes, 6*(2), 133-144.

Adams, M. L. (1980). Five co-occurring factors in speaking proficiency. In J. R. Firth (Ed.), *Measuring spoken language proficiency* (pp. 1–6). Washington, DC: Georgetown University Press.

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

Bachman, L. (1991). What does language testing have to offer? *TESOL Quarterly*, *25*(4), 671-704.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

Bailey, K. (1983). Foreign teaching assistants at U.S. universities: Problems in interaction and communication. *TESOL Quarterly*, 17, 308-310.

Bailey, K. (1984). The “foreign TA problem”. In K. Bailey, F. Pialorsi, & J. Zukowski/Faust (Eds.), *Foreign teaching assistants in U.S. universities* (pp. 3-15). Washington, D.C.: National Association for Foreign Student Affairs (NAFSA).

Boersma, P., & Weenink, D. (2015). Praat (Version 6.0.06) [Computer software]. Retrieved from http://www.praat.org.

Bowden, H. W. (2016). Assessing second-language oral proficiency for research. *Studies in Second Language Acquisition*, 1-29. doi: 10.1017/S0272263115000443

Bresnahan, M. J., & Cai, D. H. (2000). From other side of the desk: Conversations with international students about teaching in the U.S. *Communication Quarterly, 48*(2), 65-75.

Brooks, L. & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT^{TM} and real-life academic speaking activities, *Language Assessment Quarterly, 11*(4), 353-373, doi: 10.1080/15434303.2014.947532

Bulté, B. (2007). Measure for measure: Why type/token ratio based measures are not valid to assess lexical complexity/richness as a dimension of language proficiency. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 27-35). Brussels: KVAB.

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.

Bulté, B., & Housen, A. (2015). Evaluating short-term changes in L2 complexity development. *Círculo de Lingüística Aplicada a la Comunicación, 63*, 42-76. doi: 10.5209/rev_CLAC.2015.v63.50169

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication*. New York, NY: Longman.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics, 6*(2), 5-35.

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

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple... *Language Testing*, *29*(1), 19. doi: 10.1177/0265532211417211

Chiang, S. (2009). Dealing with communication problems in the instructional interactions between international teaching assistants and American college students. *Language and Education*, *23*(5), 461-478. doi: 10.1080/09500780902822959

Choi, I. (2017). Empirical profiles of academic oral English proficiency from an international teaching assistant screening test. *Language Testing*, *34*(1), 49-82. doi: 10.1177/0265532215601881

Cohen, A.D., 1998: *Strategies in learning and using a second language*. Essex, U.K: Longman.

Council of Graduate Schools (2013, November 5). First-time enrollment of international graduate students up 10 percent. Retrieved from http://www.cgsnet.org/first-time-enrollment-international-graduate-students-10-percent.

Cumming, G. (2009). Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in Medicine*, *28*(2), 205-220. doi:10.1002/sim.3471

de Jong, J. H. A. L., & van Ginkel, L. W. (1992). Dimensions in oral foreign language proficiency. In L. Verhoeven & J. H. A. L. de Jong (Eds.), *The construct of language proficiency* (pp.187–205). Amsterdam: John Benjamins.

de Jong, N. H. (2013). *Analysis of fluency*. [PowerPoint Slides, LANGSNAP Workshop, University of Southampton]. Retrieved from http://langsnap.soton.ac.uk/linked_files/LANGSNAP_dejong.pdf

de Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, *34*(5), 893-

916. doi: 10.1017/S0142716412000069

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. doi:10.1017/S0142716413000210

de Jong, N., & Vercellotti, M. L. (2015). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research*, 1-18. doi: 10.1177/1362168815606161

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, *41*(2), 385-390. doi:10.3758/BRM.41.2.385

Dörnyei, Z., & Scott, M. L. (1997). Communication strategies in a second language: Definitions and taxonomies. *Language Learning, 47*, 173–210.

Douglas, D., & Selinker, L. (1992). Analyzing oral proficiency test performance in general and specific purpose contexts. *System*, *20*(3), 317-328.

Elder, C. (1993). Language proficiency as predictor of performance in teacher education. *Melbourne Papers in Language Testing, 2*(1), 1–17.

Elder, C. (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing*, *18*(2), 149-170. doi: 10.1177/026553220101800203

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

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

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.

Faerch, C., & Kasper, G. (1983). (Eds.). *Strategies in interlanguage communication*. London, UK: Longman.

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology, 11*(1), 13-22. doi: 10.1027/1614-2241/a000086

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3), 354-375. doi: 10.1093/applin/21.3.354.

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. doi: 10.1017/S0267190515000082

Fulcher, G. (2015). Assessing second language speaking. *Language Teaching, 48*(2), 198-216. doi:10.1017/S0261444814000391

Gorsuch, G. J. (2011). Improving speaking fluency for international teaching assistants by increasing input. *TESL-EJ, 14*(4), 1-25.

Gujarati, D. (2015). *Econometrics by example* (2nd ed.). London, UK: Palgrave Macmillan.

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

Halleck, G. B., & Moder, C. L. (1995). Testing language and teaching skills of International Teaching Assistants: The limits of compensatory strategies. *TESOL Quarterly, 29*(4), 733-758.

Halleck, G. B. (2008). The ITA problem: A ready-to-use simulation. *Simulation & Gaming, 39*(1), 137-146. doi: 10.1177/1046878107308060

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association, 81*(396), 1000-1004. doi: 10.2307/2289074

Hoekje, B. (2016). “Language,” “communication,” and the longing for the authentic in LSP testing. *Language Testing, 33*(2), 289-299. doi: 10.1177/0265532215607921

Hoekje, B., & Linnell, K. (1994). Authenticity in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly, 28*(1), 103-126.

Hoekje, B., & Williams, J. (1992). Communicative competence and the dilemma of international teaching assistant education. *TESOL Quarterly, 26*(2), 243-269.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461-473. doi:10.1093/applin/amp048

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

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy, and fluency: Definitions, measurements and research. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA* (pp. 1-20). Amsterdam: John Benjamins.

Hymes, D. (1972). *Foundation in sociolinguistics*. Philadelphia, PA: University of Pennsylvania Press.

Inglis, M. (1993). The communicator style measure applied to nonnative speaking teaching assistants. *International Journal of Intercultural Relations, 17*, 89-105.

Inoue, C. (2016). A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research. *The Language Learning Journal*, 1-19. doi: 10.1080/09571736.2015.1130079

Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language, *Language Assessment Quarterly, 3*(2), 151-169. doi: 10.1207/s15434311laq0302_4

Iwashita, N., Ortega, L., Rabie, S., & Norris, J. M. (2008). *Syntactic complexity and oral proficiency in crosslinguistic perspective*. Honolulu, HI: University of Hawai‘i National Language Resource Center.

Iwashita, N. (2010). Features of Oral proficiency in task performance by EFL and JFL learners. In M. T. Prior. Y. Watanabe, & S. Li (Eds.), *Selected proceedings of the 2008 Second Language Research Forum: Exploring SLA perspectives, positions and practices* (pp. 32-47). Somerville, MA: Cascadilla Proceedings Project.

Jacobs, R. A. (1995). *English syntax: A grammar for English language professionals*. New York, NY: Oxford University Press.

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. *The Modern Language Journal*, *87*(1), 90-107. doi: 10.1111/1540-4781.00180

Johnson, K. (2004). On the Systematicity of Language and Thought. *The Journal of Philosophy*, *101*(3), 111–139.

Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics*. Oxford, UK: Oxford University Press.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 579-589. doi:10.1093/applin/aml029

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

Lennon, P. (1995). Assessing short-term change in advanced oral proficiency: Problems of reliability and validity in four case studies. *ITL Review of Applied Linguistics, 109*, 75-100.

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

Levin, J. R., Serlin, R. C., & Seaman, M. A. (1994). A controlled, powerful multiple-comparison strategy for several situations. *Psychological Bulletin, 115*(1), 153

Kane, M. (2012). Articulating a validity argument. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 34-47). New York, NY: Routledge.

Kang, O., Rubin, D., & Lindemann, S. (2015). Mitigating US undergraduates' attitudes toward international teaching assistants. *TESOL Quarterly, 49*(4), 681-706. doi: 10.1002/tesq.192

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: Massachusetts Institute of Technology Press.

Lindemann, S. (2017). Variation or 'error'? Perception of pronunciation variation and implications for assessment. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment* (pp. 193-209). doi: 10.21832/ISAACS6848

Malicka, A. (2014). *The role of task complexity and task sequencing in L2 monologic oral production*. (Doctoral dissertation). Retrieved from http://www.tdx.cesca.cat/bitstream/handle/10803/285587/Aleksandra_Malicka_THESIS.pdf?sequence=1

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381-392. doi: 10.3758/BRM.42.2.381

McGregor, L. A. (2007). *An examination of comprehensibility in a high stakes oral proficiency assessment for prospective international teaching assistants* (Doctoral dissertation). Retrieved from ProQuest LCC (UMI: 3330150).

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching, 45*(3), 241-259. doi: 10.1515/iral.2007.011

Mirdamadi, F. S., & de Jong, N. H. (2015). The effect of syntactic complexity on fluency: Comparing actives and passives in L1 and L2 speech. *Second Language Research, 31*(1), 105-116. doi: 10.1177/0267658314554498

Mirshahidi, S., & Saeli, H. (2016, March). *Construct validation of ITA proficiency using measures of complexity, accuracy, and fluency*. Paper presented at Georgetown University Round Table (GURT), Washington, DC.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73–97.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*(4), 520–531.

Nakatani, Y. (2006). Developing an oral communication strategy inventory. *The Modern Language Journal, 90*(2), 151-168. doi: 10.1111/j.1540-4781.2006.00390.x

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-578. doi:10.1093/applin/amp044

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

Oxford, R. L. (2003). Language learning styles and strategies: An overview. Retrieved from http:// http://web.ntpu.edu.tw/~language/workshop/read2.pdf.

Pallotti, G. (2009). CAF: Defining, refining and differentiation constructs. *Applied Linguistics, 30*(4), 590-601. doi:10.1093/applin/amp045

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research, 31*(1), 117-134. doi: 10.1177/0267658314536435

Pan, M. (2016). *Nonverbal delivery in speaking assessment: From an argument to a rating scale formulation and validation*. Singapore: Springer Singapore.

Papajohn, D. (1999). The effect of topic variation in performance testing: The case of the chemistry TEACH test for international teaching assistants. *Language Testing, 16*(1), 52–81.

Plonsky, L., & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics, 36*, 73-97. doi: 10.1017/S0267190516000015

Rabab'ah, G. (2016). The effect of communication strategy training on the development of EFL learners' strategic competence and oral communicative ability. *Journal of Psycholinguistic Research*, *45*(3), 625-651. doi: 10.1007/s10936-015-9365-3

Révész, A., Ekiert, M., & Torgersen, E. N. (2014). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics.* doi: 10.1093/applin/amu069

Rimmer, W. (2006). Measuring grammatical complexity: The Gordian knot. *Language Testing, 23*(4), 497-519. doi: 10.1191/0265532206lt339oa

Robinson, P. (2003). Attention and memory during SLA. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 631-678). London: Blackwell Publishing

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching, 43*(1), 1-32. doi: 10.1515/iral.2005.43.1.1

Robinson, P. (2011). Second language task complexity, the cognition hypothesis, language learning and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 3-38). Amsterdam: John Benjamins.

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*(4), 533–554. doi:10.1093/applin/amp046

Saeli, H., & Mirshahidi, S. (2014, March). *Towards the development of a valid and reliable ITA performance test*. Paper presented at the American Association for Applied Linguistics (AAAL) 2014 Conference, Portland, OR.

Saif, S. (2002). A needs-based approach to the evaluation of the spoken language ability of international teaching assistants. *Canadian Journal of Applied Linguistics*, *5*(1), 145-167.

Savignon, S. J. (1983). *Communicative competence: Theory and classroom competence*. Reading, MA: Addison-Wesley.

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin, 110*(3), 577. doi: 10.1037/0033-2909.110.3.577

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

Seymour, E., & Hewitt, N. M. (2000). *Talking about leaving: Why undergraduates leave the sciences*. Boulder, CO: Westview Press.

Skehan, P. (1998*). A Cognitive Approach to Language Learning*. Oxford, UK: Oxford University Press.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*(3), 185-211. doi: 10.1177/136216889700100302

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language learning, 49*(1), 93-120.

Skehan, P. (2003). Task-based instruction. *Language Teaching, 36*, 1–14. doi:10.1017/S026144480200188X

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510-532. doi:10.1093/applin/amp047

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.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality: Complete samples. *Biometrika, 52*(3/4), 591-611. doi: 10.2307/2333709

Smith, J. A., Meyers, C. M., & Burkhalter, A. J. (1992). *Communicate: Strategies for international teaching assistants*. Englewood Cliffs, NJ: Prentice Hall.

Tamaru, Y., Yoshioka, K., & Kimura, S. (1993). A longitudinal development of sentence structures: A study of JSL adult learners. *Nihongo kyoiku: Journal of Japanese Language Teaching, 81*, 43–54.

Tarone, E. (1980). Communication strategies, a foreigner talk and repair in interlanguage. *Language Learning, 30*, 417–431. doi: 10.1111/j.1467-1770.1980.tb00326.x

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching, 54*(2), 133-150. doi: 10.1515/iral-2016-9994

Tavakoli, P., Campbell, C. and McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly, 50*(2), 447-471. doi: 10.1002/tesq.244

Tavakoli, P., & Foster, P. (2011). Task design and second language performance: The effect of narrative type on learner output. *Language Learning, 61*, 37-72. doi: 10.1111/j.1467-9922.2011.00642.x

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

Tonkyn, A. (2007). Short-term changes in complexity, accuracy, and fluency: Developing progressive-sensitive proficiency tests. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 263-283). Brussels: KVAB.

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.

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* (Doctoral dissertation). Retrieved from ProQuest LLC (UMI: 3529566).

Vercellotti, M. L. (2015). The Development of Complexity, Accuracy, and Fluency in Second Language Performance: A Longitudinal Study. *Applied Linguistics*. doi: 10.1093/applin/amv002

Vercellotti, M. L., & Packer, J. (2016). Shifting structural complexity: The production of clause types in speeches given by English for academic purposes students. *Journal of English for Academic Purposes*, *22*, 179-190. doi: 10.1016/j.jeap.2016.04.004

Witton-Davies, G. (2014). *The study of fluency and its development in monologue and dialogue*. (Doctoral dissertation). Retrieved from http://www.forex.ntu.edu.tw/en/files/writing/4092_dc0088cd.pdf.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity* (Tech. Rep. No. 17). Honolulu: National Foreign Language Resource Center.

Wood, D. (2006). Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *The Canadian Modern Language Review, 3*(1), 13-33.

Yaman, Ş. & Özcan, M. (2015). Oral communication strategies used by Turkish students learning English as a foreign language. In M. Pawlak & E. Waniek-Klimczak (Eds.), *Issues in teaching, learning and testing speaking in a second language* (pp. 142-158). Berlin: Springer.

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), 1-27.

Yule, G., & Hoffman, P. (1990). Predicting success for international teaching assistants in a U.S. university. *TESOL Quarterly, 24*(2), 227-243.

Zha, S. (2006). *The effects of a technology-supported training system on second language use strategies for international teaching assistants* (Doctoral dissertation). Retrieved from ProQuest LCC (UMI: 3253192).

Appendix A

The ITA Test rating rubric

Test Date: ………………. Rater: ……………… Testee: ……………………….

Linguistic Competence

Includes pronunciation, grammar fluency, and comprehensibility.

Very few errors, no noticeable effect on comprehensibility	14-15 ….
Few errors, occasionally affecting comprehensibility	11-13 ….
Noticeable errors, affecting comprehensibility of some keywords or phrases	7-10 …..
Many errors, often affecting comprehensibility	3-6 ….
Serious errors, barely comprehensible	1-2 ….

Interactional Competence

Aural comprehension, the ability to respond appropriately and effectively to questions, and appropriate audience awareness.

No problem in understanding, effective responses	10 ….
Generally good understanding, somewhat effective responses	8-9 ….
Some problems in understanding, some weaknesses in responses	5-7 ….
Generally weak understanding and responses	3-4 ….
Poor understanding and responses	1-2 ….

Strategic Competence

Organization of material, appropriate development of content, and effective use of strategies to compensate for linguistic weaknesses.

Well-organized material, excellent development and use of strategies	5 ….
Good organization, development, and use of strategies	4 ….
Some problems in organization, development or use of strategies	3 ….
Many problems in organization, development or use of strategies	2 ….
Serious problems in organization, development or use of strategies	1 ….

Total ………

Final Score ……….

(Multiply by 10)

Recommendations: PASS (250 or higher) PROVISIONAL PASS (240-249) FAIL (239 or lower)

Appendix B

A screenshot of a Passed ITA Test performance

Note: A written consent was obtained from the test taker for using her screenshot.

Appendix C

Sample tester's profile and coded transcript

ITA candidate's profile

Fictitious name: Rosa (female)

Degree category: Ph.D.

Program: Microbiology

L1: Farsi

ITA Test Score:

Rater 1: Linguistic competence: 14; Interactional competence: 10; Strategic competence: 5; Overall: 29
Rater 2: Linguistic competence: 12; Interactional competence: 9; Strategic competence: 4; Overall: 25
Reported ITA score: 270 [Passed]

Coding map

T-unit boundaries: //

Excluded from T-unit count: Underlined

Errors (both lexical and grammatical): Highlighted

Unintelligible word: (( ))

All-purpose word: CAPITALIZED IN GREEN

False starts: -- --

Reformulations: ….

Extract form the transcription (monologic task)

/These molecules are chemical words that the bacteria use to communicate to each other/. /When it's alone, these triangles flow away/. But when the -- when they are al -- /when they grow and they double, now there are lots of bacterial cells/. Now, the extra and -- /they are all participating in releasing these red triangles/. /Now, the extra cellular amount of these red triangles increases/ and /the bacterium cell says ‘ah/! /There are lot of red triangles/! /It means that I have lots of neighbors/; /now we’re strong enough, we can launch our attack now!’/ /There is a practical THING here for us/. /We know that we are already running out of antibiotics/. /So, we need the new kinds of antibiotic/. /But, now we know the communication system between the bacteria/. /So, if we interrupt this communication system, this can be kind of antibiotic/. /What if me make some bacteria so they cannot talk or they cannot here, so they cannot communicate to each other/. /What we can do here is to do … to make some molecules that look like the real molecule/. /I mean they are red triangles/. /We can make some molecules that look like those molecules, so they luck into the receptors of the real molecule on the surface of the bacterial cell/. Now the bacteria -- and /they jam the recognition side of the bacterial cell/. /Now the bacterial cell is deaf/. /It cannot hear STUFF anymore/. /So, they cannot communicate to each other/. /This can be kind of antibiotic/.

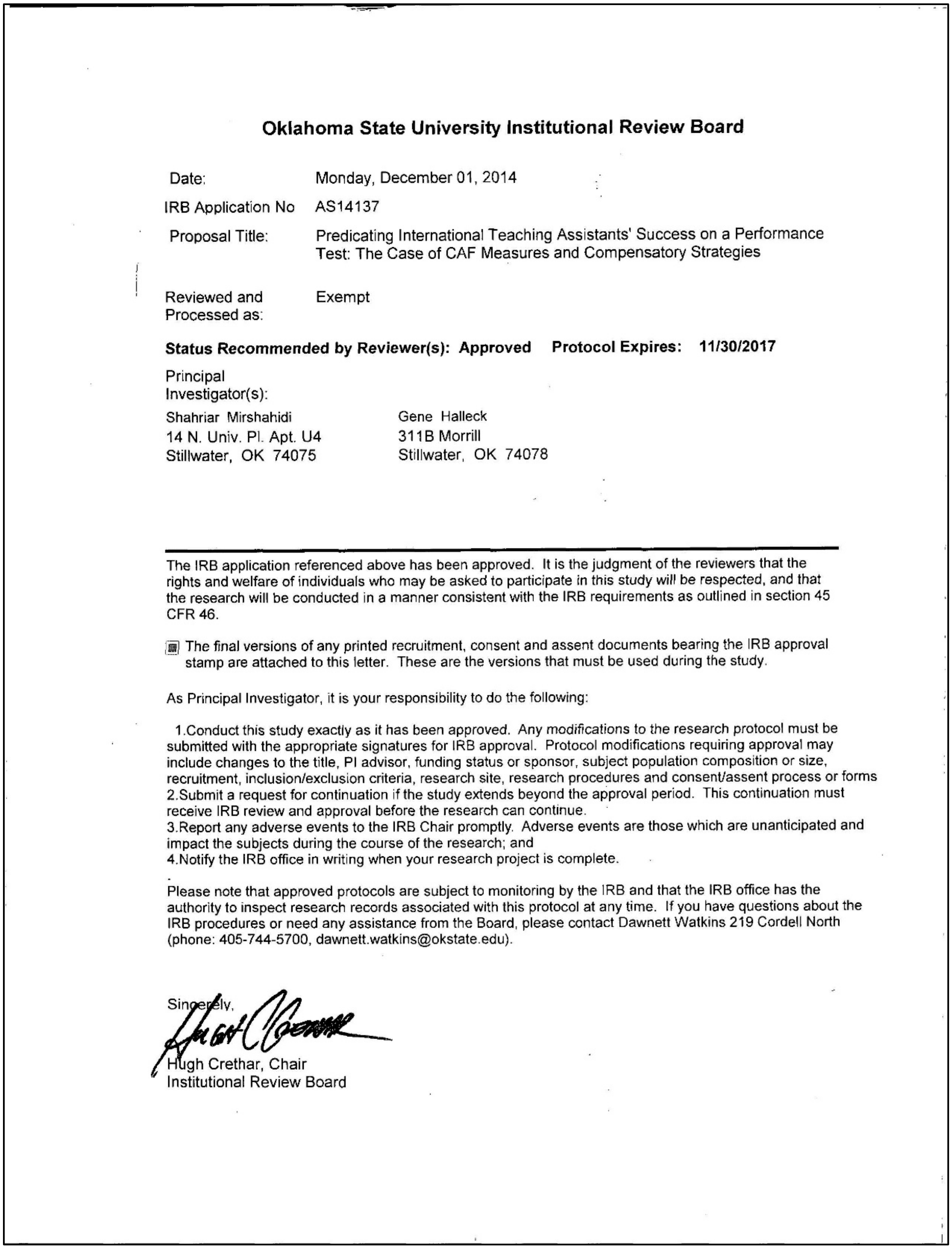
Statistics of CAF and compensatory strategy measures*

Complexity		Accuracy		Fluency				Compensatory strategies	
MLT	CL/T	PEFT	NUIW	AR	NSP	NFS	NR	APW	NLM
10.17	1.61	86.96	0	4.83	41	6	5	2	4.75

* Since two coders were involved, only the average values of the coded measures (i.e. complexity and accuracy measures as well as NFS and NR) are reported in this table. AR and NSP were analyzed in Praat based on de Jong’s (2013) script. The values represent the whole 5 minutes of the monologic performance.

Appendix D

The IRB approval letter



Oklahoma State University Institutional Review Board

Date:	Monday, December 01, 2014
IRB Application No	AS14137
Proposal Title:	Predicating International Teaching Assistants' Success on a Performance Test: The Case of CAF Measures and Compensatory Strategies
Reviewed and Processed as:	Exempt

Status Recommended by Reviewer(s): Approved Protocol Expires: 11/30/2017

Principal Investigator(s):

Shahriar Mirshahidi
14 N. Univ. Pl. Apt. U4
Stillwater, OK 74075

Gene Halleck
311B Morrill
Stillwater, OK 74078

The IRB application referenced above has been approved. It is the judgment of the reviewers that the rights and welfare of individuals who may be asked to participate in this study will be respected, and that the research will be conducted in a manner consistent with the IRB requirements as outlined in section 45 CFR 46.

☒ The final versions of any printed recruitment, consent and assent documents bearing the IRB approval stamp are attached to this letter. These are the versions that must be used during the study.

As Principal Investigator, it is your responsibility to do the following:

1.Conduct this study exactly as it has been approved. Any modifications to the research protocol must be submitted with the appropriate signatures for IRB approval. Protocol modifications requiring approval may include changes to the title, PI advisor, funding status or sponsor, subject population composition or size, recruitment, inclusion/exclusion criteria, research site, research procedures and consent/assent process or forms
2.Submit a request for continuation if the study extends beyond the approval period. This continuation must receive IRB review and approval before the research can continue.
3.Report any adverse events to the IRB Chair promptly. Adverse events are those which are unanticipated and impact the subjects during the course of the research; and
4.Notify the IRB office in writing when your research project is complete.

Please note that approved protocols are subject to monitoring by the IRB and that the IRB office has the authority to inspect research records associated with this protocol at any time. If you have questions about the IRB procedures or need any assistance from the Board, please contact Dawnett Watkins 219 Cordell North (phone: 405-744-5700, dawnett.watkins@okstate.edu).

Sincerely,

Hugh Crethar, Chair
Institutional Review Board

VITA

Shahriar Mirshahidi

Candidate for the Degree of

Doctor of Philosophy

Thesis: PREDICTING INTERNATIONAL TEACHING ASSISTANTS' PERFORMANCE IN A DOMAIN-SPECIFIC TEST: THE CASE OF COMPLEXITY, ACCURACY, FLUENCY, AND COMPENSATORY STRATEGIES

Major Field: ENGLISH (Teaching English as a Second Language)

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in TESL at Oklahoma State University, Stillwater, Oklahoma in May, 2017.

Completed the requirements for the Master of Arts in TEFL at Islamic Azad University of North Tehran, Tehran, Iran in 2010.

Completed the requirements for the Bachelor of Arts in English Translator Training at Payam Noor University, Mashhad, Iran in 2007.