Development and evaluation of online pronunciation instruction for international teaching assistants’ comprehensibility

Edna de Freitas Lima
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Bilingual, Multilingual, and Multicultural Education Commons, English Language and Literature Commons, and the Linguistics Commons

Recommended Citation
https://lib.dr.iastate.edu/etd/14561
Development and evaluation of online pronunciation instruction for international teaching assistants’ comprehensibility

by

Edna de Freitas Lima

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Applied Linguistics and Technology

Program of Study Committee:
John M. Levis, Major Professor
Carol A. Chapelle
Dan Douglas
Horabail Venkatagiri
Margaret LaWare

Iowa State University
Ames, Iowa
2015

Copyright © Edna de Freitas Lima, 2015. All rights reserved.
DEDICATION

To my husband Greg for thinking that there is nothing I cannot accomplish with my hard work and dedication and, most importantly, for making me believe it.

To my daughter Gabi for being my inspiration and for pushing me forward when the going got really tough.

To my mother Maria and to my father Paulo to whom I owe my life and whose teaching of principles and morals has shaped who I am today.

Thank you all from the bottom of my heart for your endless love, patience, and support!
# TABLE OF CONTENTS

LIST OF FIGURES ..................................................................................... v
LIST OF TABLES ....................................................................................... vi
NOMENCLATURE ..................................................................................... viii
ACKNOWLEDGEMENTS ......................................................................... ix
ABSTRACT ................................................................................................. xi

## CHAPTER 1: INTRODUCTION ................................................................. 1
  - Background ...................................................................................... 1
  - Purpose of the Study ....................................................................... 10
  - Significance of the Study ............................................................... 12
  - Outline of the Dissertation ........................................................... 13

## CHAPTER 2: LITERATURE REVIEW ..................................................... 14
  - Pronunciation Instruction and Advanced ESL Learners .................. 14
  - The Need for Careful Planning in Pronunciation Instruction .......... 15
  - Focus on Suprasegmentals .............................................................. 16
    - Word Stress ............................................................................. 17
    - Rhythm .................................................................................. 19
    - Intonation .............................................................................. 21
  - Computer-assisted Pronunciation Instruction (CAPT) ................... 22
  - Feedback and Pronunciation Instruction ....................................... 25
  - Pronunciation Feedback Through Visual Displays ....................... 27
  - Speech Awareness and Self Monitoring ....................................... 29
  - Measurement of Pronunciation Improvement ............................... 31
  - Challenges for Advanced Learners ............................................. 33
  - Principles Underlying the Design of the Supra Tutor .................... 34
  - Research Questions ..................................................................... 36
  - Chapter Summary ....................................................................... 38

## CHAPTER 3: METHODOLOGY .............................................................. 40
  - Research Design ........................................................................... 41
  - Setting .......................................................................................... 42
  - Participants .................................................................................. 43
    - Speakers ................................................................................ 43
    - Naïve Raters ........................................................................... 44
    - Trained Raters ........................................................................ 45
  - The Supra Tutor .......................................................................... 46
  - Materials ..................................................................................... 54
    - Moodle .................................................................................... 54
    - ITA Background Questionnaire ............................................. 54
    - Naïve Rater Background Questionnaire ................................ 55
    - Videotaped Lectures ................................................................ 56
    - Naïve Rater Scale .................................................................... 59
LIST OF FIGURES

Figure 1. Concentric Circles (Kilickaya, 2009, adapted from Kachru, 1997) ........................................ 2

Figure 2. World Englishes speaker-listener intelligibility matrix (Levis, 2005) ..................................... 3

Figure 3. A visual representation of the proposed study ................................................................. 12

Figure 4. Adaptation of the embedded experimental model used in this study ................................. 42

Figure 5. The Supra Tutor interface .................................................................................................. 47

Figure 6. Example of a task included in the word stress diagnostic quiz ............................................ 49

Figure 7. Instructional lecture on word stress ..................................................................................... 50

Figure 8. Tutorial on the use of online dictionaries to work on word stress ....................................... 51

Figure 9. Tutorial on how to use Praat for visual analysis of intonation patterns .............................. 52

Figure 10. Example of feedback provided on a word stress perception exercise ............................... 53

Figure 11. Screenshot of the ITA background questionnaire .............................................................. 55

Figure 12. Breakdown of the number of previous ITAs for the naïve raters ....................................... 56

Figure 13. Rating scale used by the naïve raters ................................................................................. 60

Figure 14. Rating scale used by the trained raters ................................................................................. 63

Figure 15. Screenshot of the ITAs’ evaluation of the Supra Tutor questionnaire ............................... 66

Figure 16. Speakers' perceptions of their own performance in the Supra Tutor ................................. 93

Figure 17. Speakers' evaluation of the Supra Tutor .......................................................................... 98

Figure 18. Speakers’ evaluation of the overall quality of the Supra Tutor .......................................... 98
LIST OF TABLES

Table 1. Research on the Effectiveness of Visual Feedback in L2 Pronunciation Training ........ 27
Table 2. Demographic Data of Speakers ........................................................................... 44
Table 3. Demographic Data of the Four Groups of Raters ........................................... 45
Table 4. Topics of the Pre and Post-training Lectures ................................................... 58
Table 5. Previous Research on Language Features Impacting Speaker Comprehensibility .... 62
Table 6. Summary of the Data Collection Procedures ..................................................... 71
Table 7. List of Categories and Codes for Research Question 2 ....................................... 74
Table 8. List of Categories and Codes for Research Question 3 ....................................... 75
Table 9. Inter-rater Reliability for the Data Sets Coded by the Researcher and the Data Analyst 77
Table 10. Summary of Research Questions and Data Collection Instruments .................... 78
Table 11. Data Collection Design for the Naïve Raters ..................................................... 81
Table 12. Descriptive Statistics of Pre and Post-test Comprehensibility Ratings by the Naïve Raters ..................................................................................... 82
Table 13. Independent Samples T-tests for Mean Comparison Between Pre and Post-tests ...... 84
Table 14. Descriptive Statistics for Perceived User Performance in the Supra Tutor ............ 92
Table 15. Speaker’s Favored Type of Pronunciation Instruction and Rationale ................. 96
Table 16. Descriptive Statistics for Speakers’ Evaluation of the Supra Tutor ....................... 97
Table 17. Speakers’ Rationale for their Favorite Supra Tutor Activities ............................. 100
Table 18. Data Collection Design for the Trained Raters ................................................ 103
Table 19. Descriptive Statistics of Pre and Post-test Comprehensibility Ratings by the Trained Raters ................................................................................. 105
Table 20. Raw Pre and Post Comprehensibility Ratings Assigned by the Trained Raters ..... 106
Table 21. Descriptive Statistics of Pre and Post Ratings of Suprasegmental Features ............ 107
Table 22. Frequency of Coded Suprasegmental Categories from Verbal Protocols by Proficiency Level ................................................................. 109

Table 23. Rater Comments Regarding Word Stress Issues ................................................. 110

Table 24. Frequency of Coded Suprasegmental Categories Per Rater ................................. 112

Table 25. Descriptive Statistics of Pre and Post Ratings of Non-suprasegmental Features ...... 115

Table 26. Frequency of Coded Non-Suprasegmental Categories from Verbal Protocols by Proficiency Level ................................................................. 116

Table 27. Frequency of Coded Non-Suprasegmental Categories Per Rater .......................... 117

Table 28. Rater Comments on Other Language Features ...................................................... 118
NOMENCLATURE

CAPT  Computer-assisted Pronunciation Training
CF    Corrective Feedback
EFL   English as a Foreign Language
EIL   English as an International Language
ESL   English as a Second Language
ITAs  International Teaching Assistants
L1    First Language
L2    Second Language
ACKNOWLEDGEMENTS

I would like to express my very great appreciation to my Major Professor, Dr. John M. Levis for everything he has taught me and for his infinite patience and support. Under his guidance, I have grown so much more that I could ever imagine possible. I am also very grateful to my committee members, Dr. Carol A. Chapelle, Dr. Dan Douglas, Dr. Horabail Venkatagiri, and Dr. Margaret LaWare for taking time out of their busy schedules to embark on this journey with me. Your expertise and insights have been invaluable.

I would like to offer my special thanks to Karina Silva for helping me with data analysis, but more importantly, for being with me every step of the way throughout my studies and teaching at ISU and for always giving me support and volumes of heartfelt encouragement.

I am particularly grateful to Jim Ranalli, Monica Richards, Sarah Huffman, Stephanie Link, Joe Geluso, and Jayme Wilken, my colleagues in the ISU Applied Linguistics and Technology Program, for lending their hard won expertise to the task of analyzing and rating the first ever ITAs to utilize the Supra Tutor.

I would also like to thank Mijin Jang, YunDeok Choi, Sam Ascoli, Erin Todey, Lucas Rodewald, Theresa Orlovsky, and Natalie Meyer for taking on the extra work created by allowing me to come to their English 150 classes to collect data without reservation or complaints.

In addition, I am very grateful for the effort put forth by all the English 150 students in those classes who participated in my study and applied themselves in a very professional manner to the job of rating the research participants.

I am particularly grateful for the hard work performed by the ITAs who participated in my study. As a former ITA myself, I know how little time we have to devote to taking on an
additional learning experience. All of the ITAs involved in this study spent many hours with the 
Supra Tutor. I am humbled by their willingness to sacrifice what little free time they had based 
on their faith that an untested online training program could help them to enhance their own 
teaching and learning experience. They also took the time to provide valuable feedback which I 
have taken to heart and which will improve my own ability to teach as my career progresses. 

Assistance provided by Teresa Smiley, Sheryl Kamps, and Deanna Ward was greatly 
appreciated. Sheryl and Deanna went above and beyond any reasonable expectations in helping 
me with room and equipment reservations, without which none of the data needed could have 
been collected at all. Teresa’s knowledge of graduate studies procedures and requirements 
combined with tons of deadline reminders throughout the years were essential in my academic 
success.

I will always be grateful to Iowa State University for changing my life by accepting me 
as a graduate student (twice!) and for giving me the opportunity to grow intellectually and 
professionally. Thanks to professors, mentors, program coordinators, colleagues, and staff who 
believed in me and supported me both as a student and as a teaching assistant throughout the 
years! A huge “thank you” to the hundreds of students at ISU who were a crucial part of my 
development process and who taught me as much as (if not more than) I taught them.

To my Ohio University colleagues, thank you all for your support as I completed my 
dissertation.

Finally, thank you to all of the wonderful people who have honored me with their 
friendship along the way (you know who you are). You were always there for me and made sure 
I did not flee back home when the going got tough. Thank you all!
ABSTRACT

Pronunciation is an important part of language proficiency because of its immediate role in language judgments. Nevertheless, it has often been overlooked and sometimes neglected in favor of grammar and vocabulary in SLA research (Neri, Cucchiarini, Strik, & Boves, 2002). Taking into consideration the importance of pronunciation instruction and the central impact of suprasegmentals on comprehensibility (Celce-Murcia, Brinton, Goodwin, & Griner, 2010; Isaacs, 2008), a four-week fully online pronunciation tutor (the Supra Tutor) was developed as an attempt to address gaps in pronunciation instruction identified by previous research (e.g., marginalization of pronunciation instruction, lack of trained instructors, and lack of connection between theory and practice) (Derwing & Munro, 2005).

The primary purpose of this mixed methods study, an adapted version of the embedded experimental model, was to assess whether a fully online pronunciation course can help international teaching assistants (ITAs) improve their comprehensibility. The study also analyzed ITAs’ evaluation of the Supra Tutor in regards to usefulness, level of interest, and quality of the materials. Finally, the study examined the contributions of suprasegmentals (word stress, rhythm, and intonation) to comprehensibility ratings and the language features that trained raters attended to when judging the comprehensibility of L2 speakers.

Twelve ITAs from several L1 backgrounds delivered a 7-minute field-specific lecture, which was videotaped for later analysis. Then, the ITAs completed the four-week online training. After the training, the ITAs delivered another 7-minute lecture on a related topic and were once again videotaped. Finally, the ITAs were rated for comprehensibility by different groups of naïve and trained native listeners. The rating instruments were a holistic comprehensibility scale used by naïve native listeners and a language-specific scale used by trained raters. Verbal protocols
were employed during trained rater sessions. Both scales were developed specifically for this study.

Findings indicated that the Supra Tutor was effective in providing pronunciation instruction as four out of the 12 ITAs showed significant comprehensibility improvement. The findings also indicated that the ITAs found the tutor to be useful, appealing, and of excellent quality. As for the contributions of suprasegmentals to comprehensibility, the analysis of trained rater data indicated that raters tended to place similar emphasis on rhythm and intonation. However, word stress was a suprasegmental that raters attended to when judging the comprehensibility of learners at all levels of proficiency. In addition, findings revealed that often there was not a clear connection between comprehensibility ratings and suprasegmental ratings. Also, individual trained raters focused on a variety of language features when judging L2 speech. Segmentals, especially vowels, however, were the most common source of disagreement among the raters.

The results of this study contribute to the area of technology applied to pronunciation instruction in that it is the first study to employ fully online pronunciation training through the use of a tutor designed to target specific suprasegmental features and specific learner needs. The findings in this study have several pedagogical implications. First, the tutor was perceived to keep learners engaged and motivated to complete the training, which is an indication that this kind of pronunciation instruction can be helpful to raise learners’ motivation to improve their pronunciation skills. Second, the Supra Tutor can be easily adapted for different audiences and contexts, which can help solve the lack of trained teacher conundrum. Third, in addition to learner training, an online approach such as the Supra Tutor has the potential to serve as a teacher-training tool. The Supra tutor can raise teachers’ awareness of the pronunciation features
of English, help them understand how to structure their own classes, and give them the confidence to embrace the teaching of pronunciation.
CHAPTER 1: INTRODUCTION

Background

English enjoys the distinction of having become the international lingua franca of the modern world; it is the language used by people of virtually every country to communicate in almost every conceivable context. International organizations, multinational corporations, and educational institutions all rely to a greater or lesser extent upon the ability to communicate in this global language. English has, for instance, become vital for international relations, scientific inquiry, technological advancement, international travel, and financial development (Crystal, 2012).

Kachru’s (1985) Concentric Circles (Figure 1) describe the demographics of World Englishes. The inner circle, with an estimate of up to 380 million speakers, involves all the countries identified as “native” (e.g., USA, UK, Australia, New Zealand) and is the group of countries that has exerted the most power and influence regarding linguistic standards. The outer circle includes those countries where English is used as a second language or is a nativized variety (e.g., India, Singapore, Pakistan, Jamaica). The estimated number of speakers in the outer circle ranges from 150 to 300 million. In the expanding circle, English is a foreign language rather than a second language; that is, it is not an official language, but it is often learned for international communication. The number of English speakers in the expanding circle is estimated to be up to one billion, a number larger than that of speakers in the inner and outer circles combined.
Although Kachru’s (1985) concentric circles help explain the demographics of English, the boundaries of the circles are becoming more and more blurred. In other words, the inner circle countries are becoming more similar to the outer and expanding circles as people move from one circle to another for different reasons (e.g., immigration, education, career).

The circles reflect three contexts that illustrate how oral communication in English works around the world. In the inner circle, English is used by nonnative speakers primarily for communication with native speakers. For instance, the number of non-native speakers in the United States has increased due to the number of refugees, immigrants, and individuals in search of education and/or job-related opportunities, especially in the arenas of technology, business, industry, and higher education. In the expanding circle, the purpose of learning English is to “speak the target language as a ‘foreigner’ in order to facilitate communication with NSs [native speakers] of the language” (Jenkins, 2002, p. 85). In the expanding circle, varied learners may include school children, international business employees, college and university professors, researchers, and students who wish to pursue graduate and undergraduate education in English-

**Figure 1.** Concentric Circles (Kilickaya, 2009, adapted from Kachru, 1997)
speaking universities around the world (Morley, 1991). Finally, in the outer circle, English has an official role in the administration of the country and is learned both for internal and international communication, but it may not include communication with native speakers (Jenkins, 2002). For instance, 75% of tourism takes place between non-English speaking countries, an indication of the growing use of English as an international language (Graddol, 2006).

Another way to think about how English is used throughout the world is seen in Levis’ (2005) three-by-three matrix (Figure 2) based on Kachru’s (1985) concentric circles. The nine quadrants in the matrix illustrate not only the traditional possible interactions between native speakers (NS-NS), between native and nonnative speakers (NS-NNS), and between nonnative speakers (NNS-NNS) of English in the world. They also show that outer circle speakers (neither native nor nonnative in traditional terms – Kachru calls them nativized) communicate with the other groups in ways that do not fit the traditional divide between native and nonnative speakers.

Figure 2. World Englishes speaker-listener intelligibility matrix (Levis, 2005)
Given this myriad of possible international communication scenarios, breakdowns in communication are likely to occur, especially when it comes to pronunciation. Pronunciation is an essential part of oral communication because of its immediate role in language judgments and impact on communicative competence. Non-native speakers with good pronunciation are likely to be understood even if their grammar and vocabulary are not outstanding. On the other hand, speakers with advanced grammar and vocabulary skills will not be understood if their pronunciation is below a certain threshold (Celce-Murcia & Goodwin, 1991; Hinofotis & Bailey, 1980) as poor phonetic control and prosody can distract the listener and hinder comprehension of the message (Eskenazi, 1999).

In order to have a better understanding of the importance of pronunciation in oral communication, it is necessary to define it. The term pronunciation is often used in teaching and research to encompass both segmental and suprasegmental features of English. Segmentals refer to the “inventory of sounds” (Celce-Murcia, Brinton, Goodwin, & Griner, 2010, p. 41) of English, that is, a combination of consonants and vowels. Segmentals are important because of sound distinctions, especially in regards to the functional load principle. The functional load principle, “a measure of the work which two phonemes (or a distinctive feature) do in keeping utterances apart” (King, 1967, p. 831) is proposed by Munro and Derwing (2006) as a way to prioritize issues in pronunciation teaching. The functional load principle ranks segmental contrasts in English according to their importance and their impact on speech comprehensibility. For instance, the contrastive pairs /l/-/ɾ/ (e.g., ship-sheep) and /æ/-/æ/ (e.g., bad-bed) have a relatively high functional load (i.e., the sound pairs function to distinguish a high number of words that are otherwise identical in sound), thus impacting listeners’ ability to understand a given speaker.
Suprasegmentals, on the other hand, are the features that extend beyond the individual sounds. That is, suprasegmental features (i.e., word stress, rhythm, and intonation) not only affect words but utterances, contributing to discourse structuring (Levis & Grant, 2003). In other words, suprasegmentals help listeners understand the organization of information even in the most complex English sentences (Tyler, Jefferies, & Davies, 1988). Although suprasegmentals comprise a crucial element in oral communication, they seem to be very challenging for learners to acquire. One reason for this is that learners may not be aware of the prosodic features of the target language. As a consequence, they “may have difficulties hearing, recognizing or labeling different prosodic patterns, such as segmental durations, rhythmic or intonation patterns” (Busà, 2007). If learners are not given explicit explanations of the rules governing L2 prosodic features, they are not likely to be able to make generalizations about how suprasegmentals communicate meaning in English. If prosody is taught implicitly through “listen and repeat”, it cannot be learned (Busà, 2007). Because suprasegmentals are harder to acquire and because they may have a larger impact on listener’s ability to understand than segmentals (Isaacs, 2008; Kang, 2010), they need to be explicitly taught to adult learners.

Given the importance of English for oral communication worldwide, intelligibility, the extent to which a speaker is understood (Munro & Derwing, 1995), has become pivotal in pronunciation teaching after research has shown that nativeness (native-like accuracy) is neither desirable nor attainable for most ESL and EFL learners (Levis, 2007). Pronunciation and other language skills, namely listening and speaking, are interdependent and, together, affect speech production and speech perception, facilitating and hindering communication (Gilbert, 1995; Nooteboom, 1983).
Research on the evaluation of second language (L2) speech “has indicated that non-native utterances can be evaluated along several dimensions” (Munro & Derwing, 1995, p. 290). Because there are inconsistencies in the ways in which these dimensions are interpreted and in the methods used to rate the speech of second language learners, Munro and Derwing (1995) defined three dimensions that are relevant: *intelligibility*, the extent to which an utterance is understood, *comprehensibility*, the listener’s perceptions of difficulty in understanding certain utterances, and *accentedness*, how strong the speaker’s foreign accent is perceived to be.

Smith and Nelson (1985) claim that the terms *intelligibility* and *comprehensibility* were often used interchangeably in research. In order to clarify the situation, they assigned more specific meanings to these terms and introduced a third term, *interpretability*. For them, intelligibility has to do with word/utterance recognition (listener is able to make it out and repeat it), comprehensibility concerns word/utterance meaning (locutionary force; listener can repeat it but is unable to understand the meaning), and interpretability refers to the meaning behind a word/utterance (illocutionary force; listener recognizes it but is unable to understand the intention behind it). The authors used excerpts of literary work to show that issues of comprehensibility and interpretability often happen when one is reading this genre. They argued that intelligibility, comprehensibility, and interpretability do not carry the same weight in terms of difficulty for listeners. Hence, they are not interchangeable (p. 335).

Pronunciation researchers seem to have reached the consensus that intelligibility, not nativeness (Levis, 2005), should be the aim of second language pronunciation instruction (Derwing & Munro, 2005; Hahn, 2004; Munro & Derwing, 2006) since “intelligibility is vital to successful communication” (Levis & Grant, 2003). Derwing (2010) and Isaacs (2008) add that the intelligibility principle holds that speakers should be comfortably understood. In other words,
“learners should aim to develop speaking patterns that allow them to communicate with ease, even if their accent retains nonnative characteristics” (Munro & Derwing, 2015, p. 377).

Interestingly, researchers often talk about intelligibility when, in fact, they are addressing comprehensibility. The distinction between these two dimensions is crucial for understanding the purpose of the present study. Intelligibility is primarily an issue of decoding speech (i.e., recognizing utterances) whereas comprehensibility has to do with the degree of difficulty a listener has in understanding a given message (Levis, 2006). Global intelligibility, according to Derwing (2010), is understood as comprising three types of judgments about speech, as follows:

![Intelligibility Diagram]

Intelligibility = decoding  Comprehensibility = effort  Accentedness = difference

According to Derwing’s (2010) definitions, intelligibility is the extent to which a listener actually understands a speaker, whereas comprehensibility concerns a listener’s judgment of how easy or difficult a speaker’s pronunciation is to understand. Accentedness, on the other hand, is the judgment of how one’s speech varies phonologically from the local norm (p. 29).

Research addressing the decoding of L2 speech (i.e., intelligibility) employs speech transcription (Munro & Derwing, 1995) or fixed cloze tests (Smith, 1992) as a method of assessing how intelligible a nonnative speaker is. Comprehensibility, on the other hand, is often measured through perceptual scales (Munro & Derwing, 1995; Munro & Derwing, 1998, 2001) or through other types of comprehension tests (Hahn, 2004; Williams, 1992).

As discussed above, there is confusion regarding the distinction between intelligibility (especially as a general concept) and comprehensibility, and researchers acknowledge this confusion. This study investigates how native naïve and trained listeners rate the
comprehensibility of ITAs before and after online training. Thus, from this point on, I will focus on comprehensibility, which is defined as the effort listeners have to put forth in order to understand the message conveyed by a given speaker.

As argued, pronunciation is crucial to effective oral communication and, as such, must be part of language learning and teaching. For adults, pronunciation in a foreign language, especially suprasegmentals, does not fully develop without instruction. However, effective targeted pronunciation instruction is scarce for a range of different reasons. Derwing and Munro (2005) point out that lack of support, lack of training, and difficulty in establishing goals, difficulty in defining pedagogical priorities, difficulty in deciding on effective approaches to teaching are some of the main reasons why the instruction of pronunciation is generally ignored by teachers. In addition, instructors’ heavy reliance on materials such as textbooks and software is another contributing factor to ineffective pronunciation instruction. Such materials are unlikely to fully address the needs of the learners in a particular context.

The pronunciation needs of advanced learners are not often addressed. Research shows that L2 teachers receive only limited training in phonetics or pronunciation pedagogy (Baker, 2014). This limited training drives teachers to focus mainly on controlled techniques that are less likely to transfer to free speech. This results in learners with limited ability to communicate effectively, especially in an ESL context, where they are required to interact with native speakers and show high levels of communicative competence and of comprehensibility.

One context in which comprehensibility is crucial is that of international teaching assistants (ITAs) providing instruction at American universities. The number of ITAs in the U.S. is very large. According to the Bureau of Labor Statistics, 126,030 ITAs were employed in the United States in 2014. Ninety-seven percent of these ITAs were working at colleges, universities,
and professional schools. Consequently, there is an ongoing concern about communication between undergraduate students and their ITAs both in classrooms and in office hours (Damron, 2003). ITAs play an important role at American universities as they perform a variety of instructional roles. For instance, ITAs at Iowa State University may lead recitations or discussions, work as laboratory assistants, assist professors in their courses, grade students’ assignments and tests, tutor help sessions and help rooms, or serve as the sole instructors for a course (Iowa State University: Teaching Assistant Handbook). The selection of ITAs is a high-stakes decision made by university administrators because undergraduate students tend to blame ITAs for breakdowns in communication (Lima, 2012; Plakans, 1997); therefore, it is crucial that ITAs develop high oral language proficiency, including pronunciation skills.

Research addressing native undergraduate students’ perception of ITAs indicates that these students tend to react negatively to ITAs based on different factors. One of these factors is poor language proficiency or communicative competence (Davis, 1991; Lindemann, 2002; Plakans, 1997; Rubin & Smith, 1990). Although pronunciation is only one of the factors affecting the comprehensibility of ITAs, “poor pronunciation is the most overt problem associated with ITAs, as identified by undergraduate students, language researchers, and ITAs themselves” (Isaacs, 2008, p. 560). Thus, it may be crucial that ITAs receive pronunciation instruction in order to meet oral proficiency standards.

One way to provide pronunciation instruction to ITAs is through computer-assisted pronunciation training (CAPT). With the development of technology in the last few decades, substantial research has been conducted in regards to its application to the teaching of pronunciation, which, as discussed previously, is among the most difficult skills to be acquired by adult learners learning a second language (Neri, Cucchiarini, & Strik, 2006). Studies show
evidence that CAPT, “when constructed wisely, can be both effective and flexible in addressing pronunciation instruction” (Levis, 2007, p. 185) and that it is possible to “develop CAPT environments that are realistic and pedagogically sound at the same time” (Neri, Cucchiarini, Strik, & Boves, 2002, p. 182).

There are many advantages of CAPT over traditional pronunciation instruction. Some of these advantages are individualized instruction (Levis, 2007; Seferoğlu, 2005), a private learning environment, which may reduce foreign language anxiety (Neri, Cucchiarini, & Strik, 2003, 2006; Neri, Cucchiarini, Strik, & Boves, 2002), extra learning time and material (Neri et al., 2006), possibility for self-controlled practice (Neri et al., 2002, 2006; O’Brien, 2006), flexibility, patience and additional practice time (Engwall, Wik, Beskow, & Granström, 2004), and exposure to abundant oral input (Neri et al., 2002; O’Brian, 2006).

Other advantages of CAPT include exposure to different accents, reduction in the affective filter, which is an invisible psychological filter that can either aid or impede language production in a second language (O’Brien, 2006), an authentic and appealing appearance (Neri et al., 2002), combination of videos and animations (Neri et al. 2002; Levis, 2007), and tasks aimed at developing specific skills. In sum, computer-assisted pronunciation instruction has the potential to help learners improve their oral English proficiency as it provides individualized instruction, which is not often possible in the traditional classroom.

**Purpose of the Study**

The purpose of this study was to assess the effectiveness of online pronunciation instruction on the comprehensibility of international teaching assistants (ITAs). To this end, the participants received training through a four-week fully online pronunciation tutor, The Supra Tutor, focusing on specific suprasegmental pronunciation features (namely word stress, rhythm,
and intonation) and were evaluated for comprehensibility before (pre-test) and after (post-test)
this intervention. Both the Supra Tutor and the comprehensibility scales were developed for the
study. The scales involved a holistic Likert-type comprehensibility scale used by naïve raters to
judge the overall comprehensibility of the ITAs and a Likert-type language-specific scale used
by trained raters to evaluate ITAs’ overall comprehensibility and specific language skills. The
language-specific scale was used as a tool to investigate the contributions of word stress, rhythm,
and intonation, the features targeted by the Supra Tutor, to comprehensibility ratings. Verbal
protocols were employed while trained raters rated the speakers before and after training with the
goal of shedding light on which language features native speakers attended to when judging the
comprehensibility of L2 speakers.

In addition to providing training to raise ITAs’ awareness of the suprasegmental features
of English, the tutor included an assessment component to help ITAs improve their
comprehensibility through feedback. First, the tutor included a needs analysis questionnaire to
gather background information and to gauge ITAs perceptions and training, or lack thereof,
regarding pronunciation. Based on this preliminary data, changes were made to the tutor (e.g.,
addition of instructional materials) to cater to the ITAs’ needs. The tutor also included diagnostic
quizzes, whose purpose was to allow ITAs to identify areas of weaknesses before the training.
By being aware of their shortcomings in advance, the ITAs could decide and focus on activities
that were more relevant to them. The Supra Tutor also included perception and production
feedback to help ITAs become aware of their pronunciation difficulties and develop their
comprehensibility. Figure 3 shows a visual representation of this study.
Significance of the Study

Knowledge obtained from this research will contribute to the field of applied linguistics and technology, especially in regards to pronunciation instruction. Findings from this study will have pedagogical implications to institutions, teachers, learners, and material developers in that it shows that online pronunciation instruction can be effective and a good alternative to solve the issue caused by the lack of language instructors with pronunciation training. This study also shows that individualized, targeted instruction can be achieved through an online tutor. In addition, this study has implications for language assessment in that it highlights the importance of the development of more reliable scales to fairly assess the oral proficiency of L2 speakers.
Outline of the Dissertation

This dissertation is composed of five chapters. Chapter 1 introduces the purpose and the significance of the study. The second chapter provides a literature review of topics of relevance to this study. Chapter 3 describes the methodology used in this study. It explains the research design and provides a description of the setting, of the three groups of participants, and of the materials and procedures involved in the data collection process. The fourth chapter presents the findings based on each research question. The last chapter, Chapter 5, summarizes the main findings and discusses pedagogical implications and future research directions. It also explains the limitations of this study.
CHAPTER 2: LITERATURE REVIEW

This chapter reviews research that is relevant to this study. First, it discusses the importance of pronunciation instruction to speaker comprehensibility. Next, this review focuses on the need for careful planning in pronunciation instruction given that advanced learners of English are not likely to improve their comprehensibility without explicit training. After that, this chapter establishes the importance of suprasegmentals to speech comprehensibility. Next, computer-assisted pronunciation instruction, general feedback, visual display feedback, and speech awareness and self-monitoring are explored. A discussion of how improvement is measured follows, with emphasis on why comprehensibility is an appropriate goal. Specific challenges for advanced learners, with a focus on those faced by international teaching assistants, are presented. Next, this chapter discusses the design principles guiding the development of the Supra Tutor. Finally, the research questions addressed in this dissertation are explained.

Pronunciation Instruction and Advanced ESL Learners

Pronunciation improvement is vital for the comprehensibility of spoken English. Nevertheless, it has often been overlooked and sometimes neglected in favor of grammar and vocabulary (Neri et al., 2002). It is impossible to speak without pronouncing (Luoma, 2004), and in many instances in which grammar and vocabulary are correct, below a certain level of pronunciation, communication cannot take place effectively (Celce-Murcia & Goodwin, 1991; Hinofotis & Bailey, 1980) “because poor phonetics and prosody can distract the listener and impede comprehension of the message” (Eskenazi, 1999, p. 62).

Because pronunciation is crucial for comprehensibility, pronunciation instruction is indispensable in foreign/second language learning, especially for adult learners. Advanced ESL learners typically will not improve their pronunciation beyond a certain point without explicit
instruction. It has been noted that once one reaches puberty, the ability to learn a second language, including acquiring native-like pronunciation, starts declining. Moreover, once adult learners reach a certain pronunciation plateau, their pronunciation may become “inevitably and irrevocably” fossilized (Acton, 1984, p. 71). Acton’s description is too strong, but considering that it is challenging to change pronunciation that has reached a plateau, pronunciation instruction should be given at all stages of foreign/second language learning.

A study that shows that changing fossilization is possible is that of Derwing, Munro, and Wiebe (1997). In their study, they measured the speaking improvement of 13 adult ESL learners (10 females and three males) enrolled in a twelve-week pronunciation course. They evaluated the intelligibility, comprehensibility, and accentedness of the adult learners before and after the course. The researchers noted that the learners had lived in Canada for an average of 10 years and reported having substantial interaction with native English speakers on a daily basis. According to the researchers, these adult learners would not be able to show progress over a twelve-week period without explicit pronunciation instruction. The course first focused on prosodic (suprasegmental) features and then switched focus to more localized features. Individual recordings were conducted at the beginning of the course and then again at the end of the training. Both recordings entailed the reading aloud of a list of true and false statements and were evaluated by 57 undergraduate students (monolingual speakers of English) in a linguistics course. Findings indicated that eight of the 13 learners showed an improvement in at least one of the three measures.

**The Need for Careful Planning in Pronunciation Instruction**

In order to be effective, pronunciation instruction for adult learners needs to be carefully planned. That is, targets need to be selected for maximum effect. The selection of targets should
focus on those features that have a larger impact on comprehensibility (Levis & Grant, 2003). In addition, pronunciation instruction should be learner-centered. Students’ individual needs must be taken into account. Teachers should focus on learners’ needs to improve their comprehensibility. Needs assessment is a key element in the effectiveness of pronunciation instruction. It is clear that learners from different language backgrounds will have difficulties with different pronunciation features. For instance, Japanese learners of English are likely to have difficulties distinguishing between /r/ and /l/ because the Japanese sound inventory does not include these liquid consonants (Flege, Takagi, & Mann, 1995). Arabic speakers may have a problem distinguishing between /p/ and /b/ because Arabic does not have /p/. Brazilians are likely to have a hard time pronouncing the /θ/ sound because it is not a sound in the Brazilian Portuguese repertoire. However, these are generalizations. It is important to bear in mind that even learners from the same L1 may vary in their difficulties regarding English segmentals.

Suprasegmentals, on the other hand, can be a challenge for a wider variety of learners from different L1s as suprasegmentals are harder to master (Busà, 2007). One possible explanation for this is that learners may lack awareness of the prosodic features of English. If they are not aware of these features, they are not able to recognize the prosodic patterns, let alone label them. Stress, rhythm, and intonation patterns are thus more challenging for adult learners to change (Acton, 1984). Change in these areas is not likely to occur without explicit, targeted instruction.

**Focus on Suprasegmentals**

While a command of segmental features in English is important, it may be less critical for successful communication than a command of suprasegmental features. Errors caused by mispronunciation of segmentals (e.g., ‘My son swallowed a pill’ versus ‘My son swallowed a
peel’) generally lead to minor misunderstandings that are easily repairable. Poor mastery of suprasegmentals, on the other hand, may cause more serious misunderstandings. For instance, learners’ incorrect rhythm patterns may be frustrating to native listeners, learner’s incorrect use of intonation patterns may lead learners’ to be perceived as impolite, uninvolved, or even rude, and depending on the stress patterns used, learners may not be understood at all (Celce-Murcia et al., 2010, p. 163).

Given that suprasegmentals may be more challenging to master and may have a larger impact on comprehensibility than segmentals (Celce-Murcia, Brinton, Goodwin, & Griner, 2010; Isaacs, 2008; Jenkins, 2000; Kang, 2010), they should be a central part of pronunciation teaching (Kang, 2010), especially for advanced adult learners. Researchers claim that placing focus on suprasegmentals may improve speaker comprehensibility so greater progress can be attained (Celce-Murcia, Brinton, & Goodwin, 1996). Word stress, rhythm, and intonation are features that are “suggested by the majority of phonology authorities” as having implications for speaker comprehensibility (Jenkins, 2000, p. 39). The next sections look at the relation of these features to comprehensibility.

**Word Stress**

The significance of word stress lies in its impact on understanding (Benrabah, 1997; Celce-Murcia et al., 1996, 2010; Cooper, Cutler, & Wales, 2002; Cutler & Clifton, 1984; Field, 2005; Guion, Clark, Harada, & Wayland, 2003; Isaacs & Trofimovich, 2012; Jenkins, 1998; Murphy, 2004; Tanner & Landon, 2009; Zielinski, 2008). Stress is often synonymous with rhythm in discussions of the English language. Word stress is lexical (word-level), while rhythm is phrasal (sentence-level). Benrabah (1997) argues that word stress is essential for effective communication; he argues that there are three main reasons why word stress is important for
pronunciation teaching: 1) it tends to be accompanied by syllable lengthening in the case of stressed and reduction of non-stressed syllables; 2) “the foreign learner is further confused by the lack of consensus among scholars and an absence of simple rules” (p. 159); and 3) the exact allocation of word stress can play an important role in speech processing and serves as a clue to understanding words. Benrabah (1997) argues that word stress is essential for effective communication; he also discusses the effect of stress misplacement on comprehension.

Native English-speaking listeners pay attention to various clues in identifying word stress. Guion, Clark, Harada, and Wayland (2003) present the results of experimental studies of word stress perception and production. The main aim of the study was to further the understanding of the factors influencing the placement of main stress in English by native speakers. The three factors on stress placement in two-syllable nonwords under investigation were syllabic structure, lexical class (noun versus verb), and stress pattern of words that are phonologically similar. The participants were 17 American English speakers (seven males, 10 females) with ages between 18 and 54 years. The study consisted of a production task, a perception task, and a word similarity task. The stimuli for the production task were four two-syllable word types, with 10 tokens of each syllabic type and participants were asked to concatenate the isolated syllables into a single word and utter it in a carrier phrase. Each of the syllable pairs was presented two times, once with the noun frame and the other with the verb frame. For the perception task, 40 words were produced with main stress on the first and second syllable in each of the carrier frames (“I’d like a__” and “I’d like to __”), totaling 160 sentences. Participants were also asked to listen to pre-recorded phrases in pairs and indicate which one sounded more like a real English sentence to them. The similarity task aimed at gathering information on main stress placement of words that were phonologically similar to the nonwords
used in the study. Participants were asked to indicate if the sounds reminded them of any real words in English. The results showed that all the three factors had a significant contribution to the prediction of main stress assignment. The authors concluded by arguing that the empirical results that they obtained support models of stress placement that allow multiple and potentially competing factors to play a role in stress assignment of English words.

Field (2005) analyzed the effect of manipulated lexical stress and vowel quality and their effect on both native and nonnative listeners. Results showed that both native and nonnative listeners responded in similar ways to the issues posed by stress misplacement on two-syllable words (loss of intelligibility of 19.78% for NLs and 21.28% for NNLs). Intelligibility was more affected when the stress was shifted to the right than when it was shifted to the left. Field concluded that “for both groups, the extent to which intelligibility was compromised depended greatly on the direction in which stress was shifted and whether changes in vowel quality were involved” (p. 399) and that lexical stress should be ranked at a medium level of importance for pronunciation teaching programs.

**Rhythm**

Rhythm, the second suprasegmental selected for this study, is created by the alternation of stressed syllables and non-stressed syllables (Celce-Murcia, Brinton, & Snow, 2013) in English. English is often described as a stress-timed language. Simply put, stress timed means that “the start of each stressed syllable is said to be equidistant in time from the start of the next stressed syllable” (Setter, 2006, p. 763). The importance of rhythm to comprehensibility lies in the fact that native listeners rely on rhythm in segmenting speech (Murty, Otake, & Cutler, 2007). When learners do not succeed in making the distinction between stressed and non-stressed syllables, native listeners may fail to understand the message (loss of comprehensibility).
Research shows that the use of more English native-like patterns of rhythm increases the comprehensibility of nonnative speakers. For instance, Tajima, Port, and Dalby (1997) conducted a study in which the speech of a Chinese speaker of English was synthetically manipulated to approximate a native speaker model. In order to modify the stimuli, the researchers deleted or corrected sounds (e.g., epenthetic vowels and simplified consonant clusters), shortened or lengthened segments, and added silence. The intelligibility of the original speech and of the manipulated speech was measured based on an identification task performed by native English listeners. The findings indicated that the intelligibility of the original (i.e., unmodified) speech was 39% and increased to 58% after the temporal correction. This study highlights the importance of teaching rhythm to nonnative speakers of English to improve their comprehensibility.

Hahn (2004) argues that awareness of how several prosodic features, including rhythm, affect the way native speakers process nonnative speech should affect the training of international teaching assistants. In her 2004 study, Hahn investigated the reactions of native English speakers (NESs) to nonnative primary stress in English discourse. The study involved 90 participants who were first-semester freshman students at a large Midwestern university who had no significant differences in exposure to nonnative speech. The participants were randomly assigned to three experimental groups. The stimuli were three versions of a speech given by a Korean international teaching assistant with high proficiency in English and experience as an ITA. The versions contained correctly placed primary stress, incorrectly placed primary stress, and missing primary stress. The participants had two tasks: understand and remember the content of the lecture and monitor for a tone presented sporadically (3-7 second interval) in the background of the speech and click the mouse. The computer measured the reaction time
between the tone and the mouse click. The procedures were the same for all the three versions.
The results indicated that participants significantly remembered more content of the lecture with correctly placed primary stress and that participants listening to this lecture had a shorter time reaction to the randomized tones. Hahn also found that participants tended to process this speech more easily, but this result was not statistically significant. As for misplaced and missing stress, misplaced primary stress may hinder comprehension more than missing primary stress. Hahn’s overall conclusion is that correctly placed primary stress in extended nonnative discourse facilitates communication.

**Intonation**

Intonation, the third and last suprasegmental feature selected for this study, is an important feature for comprehensibility. Correct use of intonation helps the listener understand the content of the message. For instance, by “raising pitch as one starts to speak, a speaker helps listeners orient themselves in the flow of information” (Hincks, 2005, p. 576). Intonation is “the use of suprasegmental phonetic features to convey ‘postlexical’ or sentence-level pragmatic meanings in a linguistically structured way” (Ladd, 2008, p. 4). Three primary features related to intonation are prominence, pitch range, and final intonation. Levis (2001) discusses the importance of intonational prominence in the teaching of pronunciation for conversational purposes as the role of this feature is to highlight important information in the discourse. Levis emphasizes that prominence is responsible for greater length, pitch movement, and loudness on a syllable, causing it to be more prominent than the rest of the syllables in the phrase. Levis argues that focus may be the most essential role of intonation and probably the most teachable one.

Researchers have also argued that intonation should be taught to learners at all levels of proficiency. Firth (1992) claims that at the elemental level, teachers can focus on the use of
intonation patterns. At more advanced levels of proficiency, however, the focus should shift to intonational discourse. For instance, native speakers of American English use falling final intonation for wh-questions (intonation patterns) as in “What did you buy?” but use rising intonation when using the same question signaling a request for repetition (“What did you buy?”).

Incorrect use of intonation can cause communication breakdowns between ITAs and undergraduate students. For example, by constantly using a level intonation pattern, students may perceive ITAs to be boring, unenthusiastic, or even rude. Therefore, intonation must be included in the training of ITAs. Pickering (2001) investigated the role of tone choice (the choice of a rising, falling or level pitch movement) in the communication of both international and domestic teaching assistants. In her study, she compared tone choice in the teaching demonstrations of six Chinese and six American male teaching assistants. The teaching presentations were audio and video recorded during actual lectures on chemistry, math, physics, and electrical engineering in courses solely conducted by the TAs and later analyzed for tone choices. The findings revealed that the American TAs made systematic use of tone choice to make the materials more accessible to the students and to build a connection with them. The tone choice of the Chinese TAs, on the other hand, obscured the information structure and “frequently characterized these speakers as unsympathetic and uninvolved” (p. 233). Pickering’s study highlights the importance of intonation in the teacher-student interaction and the need for training ITAs to properly use this feature.

**Computer-assisted Pronunciation Instruction (CAPT)**

In order for pronunciation instruction to be effective and have an impact on speaker comprehensibility, presentation, feedback and practice need to be consistent. However, this is
often not the case. Research has shown that pronunciation instruction has been marginalized in favor of other language skills such as grammar and vocabulary (Derwing & Munro, 2005; Neri et al., 2002). In addition, there is a lack of trained instructors and lack of support to instructors. Due to this lack of support, instructors have difficulties establishing goals and deciding on effective approaches to pronunciation instruction, which leads teachers to rely on materials that disregard learners’ individual needs (Derwing & Munro, 2005). In addition, due to lack of proper training teachers end up addressing pronunciation “unsystematically, applying it primarily as a corrective measure when errors are too prominent to be ignored” (Levis & Grant, 2003, p. 13). Because of these shortcomings in teacher preparation, computer-assisted pronunciation training (CAPT) may be a solution that provides greater consistency in training.

CAPT has the potential to address basic principles leading to effective pronunciation instruction: individualized instruction (Levis, 2007) access to multiple speech models (different speakers and different accents) (Celce-Murcia, Brinton, Goodwin, & Griner, 2010); opportunities for large amounts of practice (Celce-Murcia et al., 2010); a non-inhibiting learning environment (Eskenazi, 1999); self-determined pace (Wang & Munro, 2004); and the incorporation of technological tools that aim to facilitate and enhance the learning experience (Celce-Murcia et al., 2010).

CAPT has proven to be effective to help learners improve their pronunciation. For instance, Wang and Munro (2004) showed that computer-based perceptual training in English vowels using both synthetic and natural speech can be effectively implemented even when students have control over the amount of practice that they receive. The study tested the effectiveness of computer-based training on three vowel contrasts - /i/-/ɪ/, /u/-/ʊ/, and /ɛ/-/æ/ using an experimental and a control group. Wang and Munro’s rationale for choosing these three
contrasts was that they were common errors and that learners “tend to rely inappropriately on length as a means of distinguishing the two sounds” (p. 542) rather than vowel quality. The objective was to train learners to focus on vowel quality instead of focusing on their length. The results showed positive responses to the four posed research questions. A key finding was that the trainees’ perception improved significantly on all the three phonetic contrasts, whereas the control group did not show such improvement. Since the trainees decided their own schedule and amount of training, the authors concluded that this kind of training can be successfully applied in settings in which learners work according to their own preferences.

Tanner and Landon (2009) investigated the effectiveness of CAPT for suprasegmental improvement. Their 13-week experimental study focused on pausing, word stress, final intonation patterns, and learners’ perceived comprehensibility level. Seventy-five adult ESL learners enrolled in a university ESL program were placed into control and treatment groups. The treatment group received an overview of the targeted features to raise their awareness of prosody and was exposed to 11 weeks of self-directed computer-assisted training using Cued Pronunciation Readings, which are texts marked for a given feature (e.g., pausing). The researchers collected speech perception and production samples in the first week of the training and again in the last week. The effect of the training on the learners’ perception and production of the targeted suprasegmentals was analyzed by 10 native speakers (five males and five females) enrolled in a graduate TESOL program. The findings indicated that the treatment group made significant gains in the perception of pausing, perception of word stress, and controlled production of word stress.
Feedback and Pronunciation Instruction

Research on L2 pronunciation development emphasizes the importance of effective feedback. According to Morley (1991), the teacher’s role should be that of a “speech coach or pronunciation coach” (p. 507). A speech/ pronunciation coach provides information, provides models from time to time, offers cues, suggestions and constructive feedback about learners’ performance, sets high standards, provides a broad variety of practice opportunities, and in general, provides support and encouragement to learners (p. 507). In Morley’s (1991) view, the pronunciation “coach” facilitates learning and self-monitoring through perception and production practice.

Saito and Lyster (2012) claim that the role of teachers’ immediate feedback “might be relatively important for pronunciation teaching” (p. 627) because learners need to receive instructor’s feedback on the intelligibility of their output. In addition, learners need to practice the correct form based on a pronunciation model. Based on these two points, the authors suggest that recasts (i.e., reformulation of a student’s utterance, removing the error) may be an effective type of feedback in pronunciation development but recommend further research on other types of corrective feedback (CF) and their effectiveness.

Although pronunciation researchers highlight the importance of instructors’ feedback on the development of L2 pronunciation, there is “an absence of research specifically investigating CF effectiveness in phonological development” (Saito & Lyster, 2012, p. 598). In addition, the little empirical research available presents a few shortcomings. Three of these shortcomings are discussed below.

The first issue concerns the lack of identification of types of errors corrected. For instance, Lee (2013) investigated the types of feedback that students and teachers prefer in order
to improve students’ oral proficiency. While Lee concluded that the majority of adult learners in her study prefer explicit corrective feedback, she did not provide an account of the types of errors and mistakes corrected. The researcher also emphasized that teachers disagree that they should correct all of the errors and mistakes that students make but once again failed to discuss what type of errors and mistakes teachers tended to focus on.

Another issue is that pronunciation research tends to focus on feedback at the phoneme level (pronunciation of individual sounds) (Engwall & Bälter, 2007; Saito & Lyster, 2012) rather than on prosodic features such as rhythm and intonation. Although certain individual sounds (those which carry a high functional load – e.g., /l/-/i/) are important (Munro & Derwing, 2006), suprasegmentals are believed to have a larger impact on comprehensibility (Celce-Murcia et al., 2010; Jenkins, 2000). Therefore, it is crucial that research focus on the effectiveness of feedback on prosody to improve speaker comprehensibility.

Finally, Saito and Lyster (2012) raised the issue that the relevance of studies conducted under strict laboratory conditions “to real classrooms can only be indirect at best” (p. 597) given that these studies focus on the isolated teaching of “difficult” sound rules in which intensity and consistency of instruction are well controlled. In addition to well-controlled variables, the authors added length of instruction (in some cases, several hours on only one phonological target) to the problem of the connection between laboratory research and practical classroom application.

Research on L2 pronunciation feedback highlights the importance of the role of the instructor on pronunciation development. However, there is a lack of substantial research on the feedback given in the face-to-face classroom. What is more, to my knowledge there has been no research published on feedback in fully online pronunciation instruction. Therefore, this study
focuses on speech awareness and uses self-monitoring and the use of visual displays to provide pronunciation feedback to learners.

**Pronunciation Feedback Through Visual Displays**

For over forty years, research on the effectiveness of visual feedback on pronunciation training on segmentals and suprasegmentals has been conducted (see Table 1). Research has focused on two main approaches: 1) using native model aural input combined with visual display to aid L2 learners in improving their pronunciation skills; and 2) having learners compare their output with given visual representation to achieve maximum accuracy compared to native models (self-monitoring). Studies have revealed that visual feedback alone and visual feedback combined with aural feedback have a larger impact on the improvement of learners’ pronunciation skills than aural feedback alone (e.g., de Bot 1980, 1983).

**Table 1. Research on the Effectiveness of Visual Feedback in L2 Pronunciation Training**

<table>
<thead>
<tr>
<th>Research topic</th>
<th>Studies available in the literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impact of visual feedback on segmentals</td>
<td>Carey, 2004; Lambacher, 1999; McGowan, Ferrier, Chenausky, &amp; MacAuslan, 2001; Neri, Cucchiarini, &amp; Strik, 2008; Wang &amp; Munro, 2004</td>
</tr>
<tr>
<td>Impact of visual feedback on both segmentals and suprasegmentals</td>
<td>Seferoğlu, 2005</td>
</tr>
</tbody>
</table>

De Bot and Mailfert (1982) conducted a study to investigate whether visual feedback was effective in learning English intonation. Their study examined whether a 45-minute training
Session in perception would improve learners’ production of intonation patterns. The participants were ten Dutch and seven French English learners; a control group and an experimental group were created. The experimental group received a treatment with visual feedback on the intonation patterns in English. The results showed remarkable improvement for the experimental group and only slight improvement for the control group.

In a more recent study, Levis and Pickering (2004) used a Kay Elemetrics speech software, VisiPitch model IV 3950, to analyze intonation at the sentence level as compared to intonation at the discourse level. In their study, four native speakers of English were asked to read isolated sentences, then sentences contextualized in discourse. The findings suggest that sentence-level context affects patterns of intonation at the discourse level and point to the importance of “combining visual speech technology with a discourse-level pedagogical treatment in order to most effectively use computers in the teaching of intonation” (p. 506).

Through a mixed methods study, Hardison (2004) also investigated the effectiveness of visual display of pitch contours in real time in the improvement of prosodic features of 16 female American learners of French. In addition to investigating the effectiveness of speech technology on the improvement of French prosody (pitch, tempo, and rhythm), Hardison looked into generalization to novel sentences and segmental accuracy. The training lasted for three weeks (13 sessions of about 40 minutes each) and included practice with three sets of 30 sentences. The quantitative results showed significant improvement in prosody with generalization to segmental production and novel sentences. The qualitative part of the study, an anonymous written questionnaire given to participants, revealed learners’ perception of computer-assisted training as useful. Participants also reported becoming aware of several prosodic features of French during the training.
Given that intonation patterns comprise a major topic in research investigating the effectiveness of visual speech technology, Chun (1998) suggested four areas for the integration of intonation instruction with pitch visualization technology: visual feedback to learners, provision of authentic and abundant input, use of computers to record and analyze interactions between participants, and use of computer to track learner progress. For authentic input, Chun suggested that instructors gather examples from available corpora, pitch-track these examples, and present them to students. This suggestion is especially valuable for instructors and learners in EFL contexts, where target language input is scarce.

The research discussed in this section points to the overall effectiveness of visual speech technology in L2 pronunciation training, which is one of the goals of the online pronunciation tutor developed for this study. It has proven efficient in aiding learners to improve their pronunciation proficiency at both segmental and suprasegmental levels. What is more, research findings have shown that through the use of visual speech technology in training, learners are able to generalize the acquired knowledge to novel sentences and maintain segmental accuracy in addition to becoming aware of other prosodic features of the target language (e.g., Hardison, 2004).

**Speech Awareness and Self Monitoring**

While instructor feedback is important, it comprises just one component in L2 pronunciation development. Learners have to make instruction part of their own practice in order to improve. During interactions in the real world (e.g., opening a bank account), learners are unlikely to receive explicit feedback on their pronunciation errors and mistakes. Therefore, speech awareness and self-monitoring are crucial factors in the learning process (Morley, 1991). These two factors are the core elements in the online pronunciation tutor developed for this
study. The online tutor focused on prosodic features (namely stress, rhythm, and intonation), and its main objective is to raise participants’ (international teaching assistants) awareness of the prosodic patterns of English. Thus, a discussion of speech awareness and self-monitoring focusing on suprasegmentals follows.

Although suprasegmentals comprise an important element in second language acquisition, they seem to be very challenging for learners to acquire. One reason for this is the fact that learners may not be aware of the prosodic features of the target language. As a consequence, they “may have difficulties hearing, recognizing or labeling different prosodic patterns, such as segmental durations, rhythmic or intonation patterns” (Busà, 2007). If learners are not given explicit explanations of the rules governing L2 prosodic features, they are not likely to be able to make generalizations. If prosody is taught implicitly through “listen and repeat”, it cannot be learned (Busà, 2007).

Self-monitoring, “a process we use to direct attention and enhance metacognitive awareness of some aspect of our cognitive and behavioral functioning” (Ellis & Zimmerman, 2001, p. 205) is regarded by pronunciation researchers as an essential aspect of L2 pronunciation development (Dlaska & Krekeler, 2008). It is important that learners have the opportunity to evaluate themselves and develop self-monitoring skills for the features being practiced (Celce-Murcia et al., 2010). Language learners are better able to monitor their production of the target language when the focus is on form or on the structure of the language (Krashen, 1981). Chapelle (2001) adds that “conditions directing learners’ attention to linguistic form during tasks requiring meaningful language use are believed to be among the most important for learners’ acquisition of target language structures” (p. 47). Mastering fluency and pronunciation demands a huge amount of practice and accurate self-monitoring (Ellis & Zimmerman, 2001, p. 217).
Morley (1991) argues that self-monitoring “can begin as gentle consciousness-raising with the goal of helping students develop speech awareness, self-observation skills, and a positive attitude toward them” (p. 503). She suggests giving learners concrete suggestions for monitoring their speech, helping them develop a simple rehearsal technique (e.g., talking and listening to yourself), and helping them shift from being dependent on instructor-monitoring to becoming independent through self-monitoring. By helping learners develop speech awareness, speech self-monitoring skills, and speech adjustment strategies, teachers enable learners to improve their comprehensibility, communicability, and confidence in and outside of the classroom. The same objectives can be reached through the use of computer-assisted pronunciation instruction. As a matter of fact, CAPT has the potential to provide a wider range of options to raise learners’ speech awareness and to develop self-monitoring skills than traditional pronunciation instruction.

**Measurement of Pronunciation Improvement**

Pronunciation improvement has been measured through the level of accentedness that a speaker possesses. However, native listeners can understand the speech of even highly accented speakers (Derwing & Munro, 1997; Munro & Derwing, 2001). Accentedness, “the degree to which the pronunciation of an utterance sounds different from an expected production pattern” (Munro, Derwing, & Morton, 2006, p. 112), used to be regarded as one of the major factors hindering comprehensibility, an overall rating of how easy it is to understand a given speaker (Field, 2005). However, research has shown that comprehensibility may not be affected even when the accent is very strong (Derwing & Munro, 1997; Munro & Derwing, 1995). In 1995, Munro and Derwing investigated the effect of accent on sentence processing time; twenty native listeners rated a set of true or false statements uttered by 10 native speakers of English and 10
native speakers of Mandarin. The results showed that the utterances by Mandarin speakers took longer to process; however, the findings indicate that while the degree of comprehensibility influenced response times, the researchers found no evidence that accentedness *per se* had such an effect.

Another study on accent and processing time was conducted by Derwing and Munro (1997). In this study, the researchers looked into the relationship among accent, comprehensibility, and intelligibility. The results showed a divergence between perceived comprehensibility ratings and success in transcribing the texts; although native listeners assigned negative comprehensibility ratings to nonnative speech, they were able to successfully transcribe the texts they heard. According to the authors, this finding indicates that some accented but completely intelligible utterances may require additional effort or processing time, which drives native listeners to rate them as difficult to understand (less comprehensible).

What then causes listeners to rate perfectly intelligible speech as heavily accented? Munro and Derwing (1995) argue that even when an utterance by an L2 speaker is fully understood, accent may have an impact on communication because listeners may show prejudice against specific L2 speakers or against nonnative accents in general. The authors also claimed that various researchers have noticed a certain degree of irritation of native listeners towards L2 speakers. Dalton and Seidlhofer (1994) argued that “whether an utterance is accessible or not will be determined not only by the accuracy and clarity of the speaker’s enunciation, but also by the listener’s expectation and attitude, such as experience with, and tolerance of, low prestige or foreign accents” (p. 10). Other researchers also attributed this negative reaction toward accented speech to impatience (Lacina, 2002) and listeners’ inexperience with L2 speech (Lippi-Green, 1997). While research indicates that accent, or degree of accentedness, is attractive in talking
about nonnative speech, it may not be an effective way to measure the effectiveness of L2 speech, as accent does not always hinder comprehensibility (Isaacs & Trofimovich, 2012).

**Challenges for Advanced Learners**

As discussed previously, advanced learners may possess exceptional grammar and vocabulary skills but struggle with pronunciation. Past a certain age, learners are less likely to achieve native-like pronunciation and when they reach a certain plateau, it is challenging to change their pronunciation, especially at the suprasegmental level (Acton, 1984; Busà, 2007). Another challenge for advanced learners is to keep motivated if they believe they cannot improve their pronunciation after a given level. Motivation is a key factor in learners’ ability to improve. What is more, pronunciation improvement does not occur with instruction alone. It is key that learners make practice part of their routines and engage in meaningful practice outside of the classroom as there is evidence that varied amounts of self-directed practice can help them improve (Tanner & Landon, 2009; Wang & Munro, 2004). It is also critical that learners develop self-monitoring strategies. Self-monitoring strategies allow learners to constantly monitor their progress and make changes to their pronunciation.

One population of advanced learners that face a number of challenges is international teaching assistants (ITAs). Because ITAs often interact with native undergraduate students in and outside of classrooms, they need to show high levels of communicative competence and comprehensibility. Research on undergraduate students’ perceptions of ITAs reveals that these students tend to react negatively to ITAs for a variety of reasons. Poor language proficiency or communicative competence and poor pronunciation are among the factors influencing undergraduate students’ negative reactions to ITAs (Lindemann, 2002; Plakans, 1997). As Isaacs (2008) points out, ITAs themselves recognize that their pronunciation tend to get in the way of
successful communication with undergraduate students. Pronunciation may become a magnified issue when it comes to suprasegmentals (Gumperz, 1982) as suprasegmentals are “more clearly connected to functions of spoken English (Levis & Grant, 2003). For instance, the incorrect use of intonation for discourse structuring may lead undergraduate students to perceive ITAs to be indifferent and uninvolved (Pickering, 2001) and sometimes even ill-mannered. Since the selection of ITAs is a high-stakes decision by university administrators, it is critical that ITAs improve their oral proficiency, including their pronunciation skills.

Principles Underlying the Design of the Supra Tutor

As noted previously, a carefully designed online pronunciation tutor has the potential to address many of the challenges faced by advanced learners. First, the tutor can raise awareness of the suprasegmental features in English through appealing and cognitively-oriented materials (e.g., instructional videos). Second, an online tutor can focus on comprehensibility improvement by offering a multitude of perception and production activities focusing on meaningful tasks (i.e., tasks that are relevant to the target domain) to allow learners to practice the newly acquired knowledge. Third, an online tutor can offer flexibility of practice and schedule, allowing ITAs to perform individualized practice at their own pace, which is not often possible in the regular classroom. In other words, learners can focus on activities that are more relevant to their needs and progress quickly through activities that are less relevant. An online tutor also has the potential to help learners develop self-monitoring skills, which are vital for pronunciation improvement (Morley, 1991). Following is a brief discussion of the main design principles underlying the Supra Tutor, which draws on pedagogy and technology to offer learners an engaging and meaningful learning experience.
Because the Supra Tutor was fully online, the employment of technology was a chief design principle. The tutor was designed to incorporate available technology to enhance the learning experience and expose learners to multiple speech models and a variety of activities. First, diagnostic quizzes allowed learners to assess their understanding of a given topic before each module, and instant feedback was provided to help increase learners’ awareness of the targeted features. Next, dynamic instructional video-lectures focusing on the most relevant features of each topic exposed learners to a variety of speech models. Then, perception and production exercises that employed software such as Audacity and Praat gave learners the opportunity to practice and develop self-monitoring skills through aural and visual feedback. For example, clips from popular TV Sitcoms (e.g., The Big Bang Theory) were used in perception and production exercises to “make students aware of the multidimensional nature of spoken interaction” (Goodwin, 2005, p. 225). In addition to becoming aware of speech patterns (e.g., using intonation to convey emotions) by analyzing the video clips, learners recorded themselves imitating the model and compared input with output, thereby developing self-monitoring skills. A detailed description of the Supra Tutor’s components and activities is provided in Chapter 3.

Motivation, which can be defined as one’s desire to achieve a goal or complete a given task (Keller & Litchfield, 2002), was another principle guiding the development of the Supra Tutor. Increasing learner motivation is a key goal of computer-based instructional design (Barger & Byrd, 2011). By developing materials that are engaging, of high quality, and useful, students may be more motivated to learn. The Supra Tutor was developed to increase motivation in pronunciation learning, thereby influencing the learning context and the individual learner (Smit, 2002; Smit & Dalton, 2000).
Another principle underpinning the development of the Supra Tutor was the communicative framework for teaching pronunciation (Celce-Murcia et al., 2010, pp. 44-45). In this framework, pronunciation instruction begins with awareness raising and ends with practice that is less structured and requires learners to focus on both form and content of speech. The communicative framework is composed of five phases: 1) description and analysis (e.g., instructional lectures); 2) listening discrimination (e.g., perception exercises in which learners discriminate the feature); 3) controlled practice (e.g., exercises to highlight a given feature and raise learners’ awareness of that feature); 4) guided practice (production exercises in which learners monitor their production of a given feature); and 5) communicative practice, in which learners build fluency. By employing this framework, the Supra Tutor allows learners to build their knowledge and, ultimately, “be able to use the target language effectively for communicative purposes” (Celce-Murcia et al., 2010, p. 44).

Research Questions

This study evaluated the efficacy of an online pronunciation tutor to the comprehensibility of ITAs. An online pronunciation tutor (The Supra Tutor) focusing on word stress, rhythm, and intonation and two comprehensibility scales was developed. A detailed discussion of the online pronunciation tutor and of the scales is provided in the methods section (Chapter 3). Following are the three main research questions underpinning the study and a brief rationale for each of the research questions.

1. What is the impact of the online pronunciation tutor on the comprehensibility of ITAs?

This research question investigated whether the comprehensibility of advanced learners of English, 12 ITAs in this case, improved as a result of focused self-study using the Supra Tutor, a four-week online pronunciation tutor focusing on suprasegmentals (word stress, rhythm,
and intonation). As discussed in the literature review, computer-assisted pronunciation training has the potential to address basic principles leading to effective pronunciation instruction. For instance, it provides learners with opportunities for large amounts of practice (Celce-Murcia et al., 2010), allows them to self-determine their pace of learning (Wang & Munro, 2004), and incorporates technological tools that enhance the learning experience (Celce-Murcia et al., 2010).

The main goal of this research question was to determine whether domestic undergraduate students found the ITAs to be more comprehensible after the online pronunciation training. Due to the increasing amount of instruction provided by ITAs to undergraduate students on American campuses, these undergraduates represent the speakers’ target audience.

2. How do the ITAs evaluate the online pronunciation tutor in regards to usefulness, level of interest, and quality of the materials?

The purpose of this research question was to provide insight into the connection between improvement and motivation, as measured by ITAs’ completion of the online training. Improvement comes partly from appealing and useful materials and partly from individual factors that come into play when participants work and interact with the materials. How much did they like the Supra Tutor? How much time did they spend on it? The focus here is on the learners and how they used this resource to their own advantage. Did the ITAs do what they had to do to improve? For instance, as discussed previously, Wang and Munro (2004) found that the “unpaid” participants in their control group “willingly scheduled and attended sessions, and even reported that they enjoyed the activity” (p. 550). In doing so, those participants showed improvement, were able to transfer the gained knowledge to novel sentences, and maintained their improvement months after the training.
This research question is connected with Research Question 1, which investigated whether the comprehensibility of the ITAs improved after the training. Were motivation and the participants’ interaction with and use of the Supra Tutor related to comprehensibility ratings?

3. **What is the contribution of suprasegmental features (i.e., ratings of word stress, rhythm, and intonation) to comprehensibility ratings?**

Research Question 1 (RQ 1) focused on naïve raters’ perceptions of ITAs’ comprehensibility before and after training. One limitation of RQ 1 is that the quantitative data do not provide explanations as to what features listeners focused on when judging the speech of the ITAs. As a matter of fact, although listener perception is key to “the construct of comprehensibility, little is known about the dimensions that underlie listeners’ L2 comprehensibility judgments” (Isaacs & Trofimovich, 2012, p. 476). Research Question 3 is based on the use of the language-specific scale by the trained raters and how they rated the ITAs in terms of the language features that were included in the Supra Tutor (word stress, rhythm, and intonation) and on features that were not part of the pronunciation training (consonants, vowels, fluency, grammar, and vocabulary).

The purpose of this research question was two-fold: 1) to investigate whether the raters commented on the areas targeted by the tutor and whether these features were connected to improvement in speaker comprehensibility and 2) to analyze which language features native listeners attended to when judging L2 speech.

**Chapter Summary**

This chapter discussed literature of relevance to this study. First, the chapter discussed the importance of pronunciation instruction to speaker comprehensibility. Next, the review highlighted the need for careful planning in pronunciation instruction given that advanced
learners of English will not improve their comprehensibility without explicit pronunciation training. After that, the chapter established that suprasegmentals are vital for speaker comprehensibility. Next, the chapter focused on computer-assisted pronunciation training, general feedback, feedback through visual displays, speech awareness, and self-monitoring. A discussion of how improvement is measured followed, with emphasis on why comprehensibility is an appropriate goal to measure improvement in L2 oral proficiency. Then, specific challenges for advanced learners, especially in the case of international teaching assistants, were discussed. The chapter concluded with an explanation of the research questions underpinning this study.
CHAPTER 3: METHODOLOGY

The primary purpose of this study was to investigate whether a fully online pronunciation tutor could help ITAs improve their comprehensibility. The primary goals of the tutor were to raise ITAs’ awareness of English suprasegmental features, to help them develop self-monitoring skills, and, ultimately, to improve their comprehensibility. As discussed in Chapter 2, many English L2 learners in both ESL and EFL contexts do not receive pronunciation instruction, which leads to many learners with good grammar and writing skills, but poor pronunciation. A fully online pronunciation tutor has the potential to address basic principles involved in effective pronunciation training: access to multiple speech models (different speakers and different accents) (Celce-Murcia, Brinton, Goodwin, & Griner, 2010); opportunities for large amounts of practice (Celce-Murcia et al., 2010), which is not possible in a regular classroom; a non-inhibiting learning environment (Eskenazi, 1999); self-determined pace (Wang & Munro, 2004); and the incorporation of technological tools that aim to facilitate and enhance the learning experience (Celce-Murcia et al., 2010).

This chapter describes the data collection methods and materials used in this dissertation. First, it describes the research design employed in the study. Then, it provides a description of the setting and of the three groups of participants (speakers, naïve raters, and trained raters) involved in the study. Next, the chapter introduces the Supra Tutor, the online pronunciation tutor designed for this dissertation, and all of the materials and procedures used in the data collection process. A discussion of data analysis methods for each research question concludes the chapter.
Research Design

In this study, discrete measures of performance and holistic analyses of attitudes and perceptions were used to determine the effectiveness of the Supra Tutor, both in improving speaker comprehensibility and in the participants’ evaluation of the tutor. Therefore, this study adopted an embedded mixed method design (Creswell & Plano Clark, 2007). In this particular research design, the main data are quantitative, “and the qualitative dataset is subservient within that methodology” (p. 69).

Research Question 1 (What is the impact of the online pronunciation tutor on the comprehensibility of ITAs?) was addressed through the quantitative analysis of the data derived from the overall comprehensibility ratings from 178 domestic undergraduate students (naïve raters) enrolled in 13 sections of English 150, Critical Thinking and Communication, at Iowa State University. Data gathered from the Supra Tutor activity reports were used to support the findings yielded by the quantitative data.

The quantitative and qualitative data for the analysis of Research Question 2 (How do the ITAs evaluate the online pronunciation tutor in regards to usefulness, level of interest, and quality of the materials?) were derived from three different sources: a numeric and open-ended online questionnaire that the ITAs completed at the end of the training, emails from the participants, and a forum activity included in the online tutor. The forum activity prompted participants to write about their experience as graduate students in America and about their experience participating in the online pronunciation training.

Both quantitative and qualitative data were used to answer Research Question 3 (What is the contribution of suprasegmental features (i.e., ratings of word stress, rhythm, and intonation) to comprehensibility ratings?). The data for the analysis of this research question were derived
from the ratings by six trained raters (quantitative) and from 72 verbal protocol (e.g., Zielinski, 2008) files that were audio-recorded while the raters performed their rating of each speaker (qualitative). As seen in Figure 4, this study employed an adapted version of the embedded experimental model (Creswell & Plano Clark, 2007, p.68), a variant of the embedded design.

**Figure 4.** Adaptation of the embedded experimental model used in this study

**Setting**

As is the case of many universities in the United States, Iowa State University employs a large number of international teaching assistants. In order to determine the level of teaching responsibility to be fulfilled in their respective departments, the graduate students are required to take the Oral English Certification Test (OECT), which consists of an oral proficiency interview (OPI) and a teaching simulation test (TEACH). Depending on their performance, graduate students may be fully certified, conditionally certified, certified with restrictions, or not certified to teach. To be fully certified to teach, a graduate student must score a minimum of 230 points out of a total of 300 possible points. Those who score below 230 are required to take at least one semester of oral communication skills classes for ITAs and retake the test at the end of each semester.
Given that this is a high-stakes situation for graduate students and for their departments, the Supra Tutor was developed to help ITAs improve their comprehensibility without having to take a regular oral communication skills class. In other words, ITAs could work on activities that focus on their individual needs at their own pace to maximize instruction effectiveness. The main goals of the tutor were to raise ITAs’ awareness of the suprasegmental features of American English and to help them develop self-monitoring skills. As discussed in Chapter 2, awareness and self-monitoring skills are two aspects that play a crucial role in the comprehensibility improvement of non-native speakers of English (Morley, 1991).

**Participants**

**Speakers**

Twelve current and prospective ITAs, seven males and five females, from different fields of study and varied first language backgrounds participated in this study to simulate the diversity of accents on American campuses. The ITAs, ages ranging from 23 to 30 (average = 27), were from Brazil (1), China (4), Ecuador (1), Ethiopia (1), India (1), Iran (3), and Jordan (1). The ITAs (six MA students and six Ph.D. students) were recruited independently. On average, the participants had had English instruction for approximately 10 years, mostly in their countries of origin. Although all of the participants had been learning English for several years prior to this study, only six had had some kind of pronunciation training, mostly at the segmental level (i.e., vowels and consonants).

In order to eliminate topic as possible variable in this study, all 12 ITAs delivered lectures on basic topics in their respective fields of study for both pre and post tests. The topics for both pre and post-training lectures were closely related. Table 2 provides an overview of the demographics of the speakers.
Table 2. Demographic Data of Speakers

<table>
<thead>
<tr>
<th>N</th>
<th>Gender</th>
<th>Average age</th>
<th>Country of origin</th>
<th>L1</th>
<th>Fields of study</th>
<th>Student status</th>
</tr>
</thead>
</table>

Naïve Raters

To simulate the target audience, the naïve raters were undergraduate native speakers of American English enrolled in regular sections of English 150 (Critical Thinking and Communication). English 150 is an undergraduate writing foundation course focusing on the development of critical reading and thinking skills applied to topics of civic and cultural importance. The course also provides an introduction to oral, visual, and electronic communication. Research has employed undergraduate students as naïve raters (e.g., Isaacs & Trofimovich, 2012; Kang, 2010) because they represent the target audience of ITAs and because these raters provide a more realistic view of how comprehensible the speakers are. The data collection process took place during actual English 150 class time (13 sections). A total of 254 students participated in the activity. Data from seventy-six participants were excluded from the analysis. Sixty-two were international students, three did not sign the consent form, one could not play all the videotaped lectures, one was late to class and missed the background questionnaire, one student was texting on his cell phone while watching the lectures, and eight
participants had incomplete rating sheets. Data from 178 naïve raters were analyzed in this study. One hundred and twenty of these participants were males and 58 were females with ages ranging from 18 to 27 (average age = 18.49). The raters were divided into four groups so that each group watched six lectures in the allotted data collection time (50 minutes per class). Table 3 details the composition of each group. In order to eliminate a sequencing effect, each group rated six lectures (three pre-training and three post-training videos). The videos were alternated (e.g., pre, post, pre) and assigned to each group in such a way as to avoid raters rating the pre and post-training lectures of the same speaker.

**Table 3.** Demographic Data of the Four Groups of Raters

<table>
<thead>
<tr>
<th>Groups of raters</th>
<th>N</th>
<th>Gender</th>
<th>Average age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>42</td>
<td>29 males, 13 females</td>
<td>18.5</td>
</tr>
<tr>
<td>Group 2</td>
<td>38</td>
<td>25 males, 13 females</td>
<td>18.4</td>
</tr>
<tr>
<td>Group 3</td>
<td>42</td>
<td>30 males, 12 females</td>
<td>18.4</td>
</tr>
<tr>
<td>Group 4</td>
<td>56</td>
<td>36 males, 20 females</td>
<td>18.7</td>
</tr>
</tbody>
</table>

**Trained Raters**

Six trained raters, two males and four females, were used to address Research Question 3. The raters were native speakers of American English with basic linguistic training (an MA in Teaching English as a Second Language/Applied Linguistics and/or a Ph.D in Applied Linguistics and Technology) and language teaching experience. All six raters had experience teaching English to non-native speakers in the United States (ESL) and three also had experience teaching English in other countries (EFL). It was critical that the trained raters had prior linguistic training to ensure familiarity with the linguistic features that they were asked to rate during the verbal protocols. That is, trained raters are better able to express linguistic influences on their judgment of L2 speech than naïve raters (Isaacs & Thomson, 2009; Isaacs & Trofimovich, 2012). The trained raters were divided into two groups (three raters in each group).
Each group rated 12 videotaped lectures (six pre and six post-training videos) in alternated order (e.g., post, pre, post). To avoid a sequencing effect, none of the raters rated pre and post lectures from the same speaker.

**The Supra Tutor**

The Supra Tutor (Figure 5), a fully online pronunciation course, was available to participants in Moodle, a course management system widely used by the English Department at Iowa State University. The main goals of the online tutor were as follows: 1) to improve speaker comprehensibility; 2) to raise awareness of suprasegmental features of American English; 3) to promote self-monitoring skills; and 4) to transfer gained knowledge to novel sentences and contexts.

The tutor consisted of four modules: Word Stress, Rhythm, Intonation, and Review. The entire tutor included academic and field-specific vocabulary so that the ITAs could use what they learned in the target language domain (i.e., academic settings). The academic vocabulary used in the tutor came from the *Academic Word List* (Coxhead, 2000) and the field-specific vocabulary came from the book *Communicate: Strategies for International Teaching Assistants* (Smith et al., 1992).

As seen in Figure 5, the modules in the online tutor were color-coded; that is, the design of the activities within a given module included the color of each module as displayed in the tutor’s interface. Including a color scheme made the course visually appealing and engaging and helped ITAs keep track of which module they were working on.
Figure 5. The Supra Tutor interface
The tutor was designed for one week per module. Each of the first three modules focused on a given suprasegmental feature (word stress, rhythm, and intonation); the fourth module gave the participants the opportunity to revisit the topics and to do practical exercises to analyze how they did in learning the content presented in the materials. The tutor was completely online so that ITAs could work at their own pace, and a teacher was not required. Participants decided their own schedules; however, in order to control the experiment and prevent participants from completing all four modules in one single week, they were asked to work on a given feature per week. As a matter of fact, the modules were “hidden” from the participants, and each module was made available on the scheduled date and time for that module to begin. Each module included diagnostic quizzes, instructional lectures and materials, tutorials on how to use Audacity and Praat for visual feedback, and perception and production exercises. Following is a brief description and rationale for each type of activity included in the Supra Tutor.

**Diagnostic Quizzes**

The diagnostic quizzes were created using the “Quiz” feature in Moodle to allow for scoring and instant feedback. Only perception tasks were included in the diagnostic quizzes, which contained between 15 and 20 questions each. The purpose of the quizzes was two-fold: to allow ITAs to assess their previous knowledge on each suprasegmental feature and to identify areas of weaknesses and address them. That is, based on their score on the quizzes, ITAs could select and spend more time on activities that were more relevant to them. Although the tasks included in the diagnostic quizzes were perception tasks, they were designed to raise ITAs’ awareness of the targeted pronunciation features. The feedback built into the quizzes served to further help ITAs improve their awareness through brief explanations for incorrect options. Figure 6 shows an example of a task included in the word stress diagnostic quiz.
Instructional Lectures and Materials

Most of the instructional lectures were created specifically for the Supra Tutor using *PowerPoint* and *Camtasia Studio 8*, hosted on YouTube, and embedded in Moodle to avoid technology glitches. By using *Camtasia* to record and edit the lectures it was possible to make them more dynamic through the use of features to highlight information (e.g., zoom in and callouts) and transitions. The lectures were relatively short (no longer than 11 minutes each) and focused on the most relevant aspects of each suprasegmental feature. For instance, the lecture “Acoustics of Stress” (Figure 7) focused on characteristics of word stress, levels of stress in English, and visual representations of word stress using *Audacity*. In order to expose ITAs to a variety of speaker models, a few lectures available on YouTube were also incorporated into the tutor.


**Figure 7.** Instructional lecture on word stress

The instructional materials also included readings to help ITAs expand their understanding of each suprasegmental feature. Another instructional resource that was included in the tutor was a “useful tip” tutorial on how to use the Merriam-Webster online dictionary to identify where the primary (and secondary) stress falls in a given word in English (Figure 8). The tutorial, created using the software FastStone Capture, also emphasized the importance of listening to a word in the dictionary to pronounce that word correctly.
Tutorials on Audacity and Praat

Audacity, a free sound editing software, was used throughout the online tutor as a tool for ITAs to record and edit sound files for visual speech analysis (e.g., placement of primary stress and linking) and for production exercises. Praat, a free platform for speech analysis, was employed for visual analysis of speech contours (intonation patterns). The ITAs recorded and edited audio files in Audacity and used Praat to analyze whether their use of intonation patterns was appropriate. Because these two pieces of software were crucial for visual speech feedback in the Supra Tutor, tutorials on how to download, install, and use them were developed. The
tutorials were created using the screen capture feature of Camtasia so that the ITAs could perform the steps presented in the tutorials as they watched the videos. Figure 9 shows a screenshot of the tutorial on how to use Praat for visual analysis of speech contours.

**Tutorial 2: Using Praat to analyze your intonation**

![Tutorial 2: Using Praat to analyze speech](image)

**Figure 9.** Tutorial on how to use Praat for visual analysis of intonation patterns

**Perception Exercises**

The perception exercises integrated in the tutor provided ITAs with the opportunity to check their understanding of the content presented in the instructional materials and to practice. The exercises incorporated both reading and listening tasks and provided immediate feedback, which included a brief explanation for why a given option was incorrect (Figure 10). For the exercises that included listening tasks, a variety of speakers was used to expose ITAs to different voices and accents.
Figure 10. Example of feedback provided on a word stress perception exercise

Production Exercises

The production exercises comprised the last activity in each module. The ITAs performed the production activities after taking the diagnostic quizzes, watching the instructional lectures, and doing the perception exercises, and monitoring their progress throughout the modules. The production activities encompassed a variety of tasks for multi-faceted practice. For instance, ITAs were prompted to imitate a model (e.g., sitcom scenes), choose an answer from a list of options and record their answers in Audacity, role-play monologues and dialogues, sing along, or answer open-ended questions. Short American sitcom scenes were used throughout the tutor for warm-up and production activities. These sitcom scenes comprised an excellent source of authentic and appealing texts and of a variety of speaker models. The sitcoms were selected based on popularity (online reviews) and on content and context appropriateness.

The Supra Tutor also included a questionnaire at the end of the four modules to assess ITAs’ perceptions of the tutor. The information (both numeric and open-ended) gathered in the questionnaire was used in the quantitative and qualitative data analyses to provide answers to
Research Question 2 (How do the ITAs assess the Online Pronunciation Tutor in regards to usefulness, level of interest, and quality of the materials?)

**Materials**

**Moodle**

Part of the data collection process (background questionnaire, training, material evaluation questionnaire, and rating by naïve raters) for this study was conducted online. These materials were available in Moodle, an open-source course management system broadly used by the English Department at Iowa State University. The platform offers users a range of features, including questionnaires, quizzes, and surveys.

**ITA Background Questionnaire**

The main purpose of the ITA background questionnaire (Appendix A) was to serve as a needs assessment of participants’ prior knowledge, or lack thereof, of English pronunciation so that changes could be made to the tutor to cater to the ITAs’ needs. The questionnaire was designed using the “questionnaire” feature available in Moodle. It was also administered in Moodle. It consisted of demographic (e.g., name, age, country of origin) and background (e.g., length of English instruction, amount of daily use of English) information. Participants were also asked to indicate whether they had had pronunciation instruction prior to this study and to describe a situation in which their pronunciation was the cause of a breakdown in communication. A screenshot of the background questionnaire is shown in Figure 11.
Naïve Rater Background Questionnaire

As mentioned previously, the data were collected during 13 actual English 150 classes. That is, it was not feasible to select participants prior to the activity or to assign participants to a pre-defined group. Since a number of international students were also enrolled in each of those sections, the background questionnaire (Appendix B) was designed to include both populations and to facilitate the exclusion of non-native data. The questionnaire consisted of 11 questions and asked participants to provide their assigned ID number, age, gender, country of origin (if not American), native language, educational background, languages other than L1 and/or English, current major, length of residence in the U.S. (if not a domestic student), and the number of international teaching assistants that they had had prior to this study. This last question was
important to gauge how familiar students were with international teaching assistants. As seen in Figure 12, out of the 178 domestic students participating in this study, only 36 had not received instruction from an ITA, 57 had had one ITA, 55 had had two ITAs, 18 had had three ITAs, nine had had four ITAs, one student had received instruction from five ITAs, and two students had had six ITAs prior to this study.

![Number of previous ITAs for naïve raters (N = 178)](image)

**Figure 12.** Breakdown of the number of previous ITAs for the naïve raters

**Videotaped Lectures**

The stimuli for this study were 24 videotaped field-specific lectures, each lasting seven minutes. Twelve of the lectures were recorded before the training, and twelve were recorded afterwards. All 12 ITAs delivered a lecture on their specific area of study before the training and another lecture on a different, but related, topic after the training (see Table 4). In order to avoid anxiety as a possible variable, participants were asked to suggest topics that they would be comfortable teaching and that would be easy for undergraduate students from a variety of majors.
to understand. After topics were suggested, the researcher selected the materials for the participants. The lectures were recorded using a Canon Vixia HF M500 Camcorder and a Sunpak 6601 UT tripod and edited using Camtasia Studio 8 to later be evaluated for comprehensibility. The lectures were delivered in a classroom to simulate authenticity of task and context. In addition, this task is similar to that of the TEACH portion of the Oral English Certification Test (OECT), with which the ITAs were already familiar. It is important that the task is relevant in the target language use (TLU) domain (Bachman & Palmer, 2010). The target language domain in this case is the classroom, in which interactions between ITAs and students take place on a regular basis. The ITAs were given two minutes to write on the board before delivering each lecture; however, the use of other visual aids (e.g., PowerPoint slides) was avoided for two reasons: 1) they might facilitate listeners’ comprehension of the lecture and 2) they might divert raters’ attention from the oral text.
### Table 4. Topics of the Pre and Post-training Lectures

<table>
<thead>
<tr>
<th>Pre-training lecture topic</th>
<th>Post-training lecture topic</th>
<th>Main source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematical models</td>
<td>Solving mathematical models graphically</td>
<td>Winston, 2004</td>
</tr>
<tr>
<td>Introduction to Chemical Engineering</td>
<td>ChE 210 and ChE 356: Two required courses in Chemical Engineering</td>
<td>Denn, 2011</td>
</tr>
<tr>
<td>Models of heat transfer</td>
<td>Convection heat transfer</td>
<td>Bergman, Lavine, Incropera, &amp; DeWitt, 2011</td>
</tr>
<tr>
<td>Data warehousing in the age of big data</td>
<td>Big data case study: Facebook</td>
<td>Krishnan, 2013</td>
</tr>
<tr>
<td>Introduction to thermochemical processing of biomass</td>
<td>Thermochemical processing: Pyrolysis</td>
<td>Brown, 2011</td>
</tr>
<tr>
<td>An overview of journalistic ethics</td>
<td>Media ethics: Honesty</td>
<td>Jennings, 2008</td>
</tr>
<tr>
<td>Critical chain project management: The work breakdown structure</td>
<td>Milestone sequencing &amp; work packages</td>
<td>Leach, 2014</td>
</tr>
<tr>
<td>Fat digestion and absorption Part 2</td>
<td>Fat digestion and absorption The role of fungi in plant diseases</td>
<td>Christophe &amp; DeVriese, 2000</td>
</tr>
<tr>
<td>Introduction to plant pathology: Types of plant diseases</td>
<td>The role of fungi in plant diseases</td>
<td>Agrios, 1997</td>
</tr>
<tr>
<td>Alternating current and voltage</td>
<td>Resistors, inductors and capacitors</td>
<td>Kirtley, 2010</td>
</tr>
<tr>
<td>Introduction to operating systems</td>
<td>The operating system as a resource manager</td>
<td>Tanenbaum, 2009</td>
</tr>
</tbody>
</table>
Naïve Rater Scale

This study employed a 9-point Likert scale (Derwing & Munro, 2001; Munro & Derwing, 1995, 1998, 2001), with low numbers representing positive ratings (e.g., extremely easy to understand) and high numbers representing negative ratings (e.g., impossible to understand) for measuring comprehensibility. One advantage of employing this type of scale is its versatility. It can be used to evaluate the speech of English learners from any first language background at any level of proficiency on any given task (Isaacs & Thomson, 2013). This scale (Figure 13) was used by the naïve raters (see Appendix C for scale and instructions to raters). The purpose of this scale was to measure raters’ holistic perceptions of speakers’ comprehensibility, speech speed, accent, fluency, and presentation skills before and after the intervention (i.e., online pronunciation training). Although this scale included five items, only ratings of comprehensibility (Item 1) were analyzed for this study. The other four items (speech speed, accent, fluency, and presentation skills) were included so that the activity (rating of the academic lectures) fit the English 150 curriculum in regards to effectiveness of oral communication. This scale was available online because data collection from naïve raters took place in Moodle, where the videotaped lectures were hosted.
1. The speaker is

Extremely easy to understand

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impossible to understand</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. The speaker speaks

Quickly

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very slowly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. The speaker has

A native speaker accent

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>A very strong accent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. The speaker is

Extremely fluent

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely disfluent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. The speaker is

A very good presenter

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>A very poor presenter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 13.** Rating scale used by the naïve raters

**Trained Rater Scale**

While the naïve raters used a 9-point holistic scale (Figure 13), a more specific scale (Figure 14) was developed for the trained raters. This language-specific scale was designed to serve two purposes: 1) to investigate how the trained raters rated ITAs in regards to word stress, rhythm, and intonation, the three features targeted by the Supra Tutor and 2) to determine which linguistic factors had a larger impact on speaker comprehensibility according to the trained raters. Verbal protocols were employed to provide data that shed light on decisions that raters made when rating the comprehensibility of L2 speakers. In addition to overall comprehensibility (item 12 on the scale), this scale included specific linguistic features: segmentals (consonants and vowels - items 1 and 2, respectively), word stress (item 3), rhythm (stressed and unstressed words/syllables - items 4 and 5), intonation (pitch and final intonation - items 6 and 7), fluency (pauses and hesitations - item 8), grammar (item 9), and vocabulary (vocabulary range and word choice and expression - items 10 and 11). It is noteworthy that overall comprehensibility (how
easy or difficult a speaker is for listeners to understand) was rated based on Item 12. The other 11 items, or categories, were included in an attempt to investigate which language features trained raters focus on to arrive at their overall comprehensibility judgments. In other words, the other categories were meant to help deconstruct why the raters rated a speaker’s comprehensibility in the way they did. The justification for the categories and features included in the trained rater scale is based on previous research on speaker comprehensibility (see Table 5). In addition to the pronunciation features (vowels, consonants, word stress, rhythm, and intonation) discussed in Chapter 2 of this dissertation, research has shown that various other linguistic features have an impact on how understandable native listeners find L2 speakers to be. For instance, Isaacs and Trofimovich (2012) found that in addition to pronunciation (both segmentals and suprasegmentals), grammar, vocabulary, and fluency all helped distinguish levels of comprehensibility. They also found that word stress errors discriminated speakers at all levels of proficiency. Appendix D includes the scale and instructions to the trained raters.
Table 5. Previous Research on Language Features Impacting Speaker Comprehensibility

<table>
<thead>
<tr>
<th>Categories</th>
<th>Features</th>
<th>Previous Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segmentals</td>
<td>Consonants</td>
<td>Derwing, Munro, &amp; Wiebe, 1998; Gallego, 1990; Isaacs, 2008; Isaacs &amp; Trofimovich, 2012; Jun &amp; Li, 2009</td>
</tr>
<tr>
<td></td>
<td>Vowels</td>
<td>Derwing, Munro, &amp; Wiebe, 1998; Gallego, 1990; Isaacs, 2008; Isaacs &amp; Trofimovich, 2012; Jun &amp; Li, 2009</td>
</tr>
<tr>
<td>Rhythm</td>
<td>Sentence stress</td>
<td>Anderson-Hsieh, Johnson, &amp; Koeler, 1992; Derwing, Munro, &amp; Wiebe, 1998; Hahn, 2004; Isaacs, 2008; Kang, 2010; Tajima, Port, &amp; Dalby, 1997</td>
</tr>
<tr>
<td></td>
<td>Linking</td>
<td>Alameen, 2014; Jun &amp; Li, 2009</td>
</tr>
<tr>
<td>Fluency</td>
<td>Tempo and flow</td>
<td>Derwing, Munro, &amp; Wiebe, 1998</td>
</tr>
<tr>
<td></td>
<td>Flow of speech</td>
<td>Gallego, 1990</td>
</tr>
<tr>
<td>Grammar</td>
<td></td>
<td>Gallego, 1990; Isaacs &amp; Trofimovich, 2012</td>
</tr>
<tr>
<td>Vocabulary</td>
<td></td>
<td>Gallego, 1990; Isaacs &amp; Trofimovich, 2012</td>
</tr>
</tbody>
</table>
1. Consonants are always

| Clearly enunciated | 0 1 2 3 4 5 6 7 8 9 | Poorly enunciated |

2. Vowels are always

| Clearly enunciated | 0 1 2 3 4 5 6 7 8 9 | Poorly enunciated |

3. Main stress on multisyllabic words is

| Always placed correctly | 0 1 2 3 4 5 6 7 8 9 | Often misplaced |

4. Stressed words in a sentence are

| Always clearly pronounced | 0 1 2 3 4 5 6 7 8 9 | Rarely pronounced clearly |

5. Unstressed words/syllables are

| Usually de-emphasized | 0 1 2 3 4 5 6 7 8 9 | Rarely de-emphasized |

6. The voice range is

| Lively | 0 1 2 3 4 5 6 7 8 9 | Monotone |

7. Final intonation

| Always sounds natural | 0 1 2 3 4 5 6 7 8 9 | Rarely sounds natural |

8. Unnatural hesitations and pauses

| Rarely occur | 0 1 2 3 4 5 6 7 8 9 | Occur very often |

9. The use of grammar is

| Always correct | 0 1 2 3 4 5 6 7 8 9 | Rarely correct |

10. Vocabulary range is

| Sufficient | 0 1 2 3 4 5 6 7 8 9 | Very limited |

11. Word choice and expression are

| Always appropriate | 0 1 2 3 4 5 6 7 8 9 | Often inappropriate |

12. Overall, the speaker is

| Easy to understand | 0 1 2 3 4 5 6 7 8 9 | Difficult to understand |

**Figure 14.** Rating scale used by the trained raters
Although it is beyond the scope of this study to investigate the validity of the rating scale, this specific comprehensibility scale was developed taking into consideration Fulcher’s (1996) claim that “until test researchers and developers take seriously the validity of tests at the development phase rather than as a post hoc notion, the problem of the indeterminacy of validation studies and the uninterpretability of test scores will remain serious” (p. 228). That is, in developing the rating scale for this study, the researcher attempted to integrate applied linguistics, second language acquisition, and language testing theory and practice as suggested by Fulcher (1996, p. 228).

Research on pronunciation assessment has emphasized several shortcomings of the assessment scales of standardized English tests such as IELTS and TOEFL iBT (Harding, 2012; Isaacs & Trofimovich, 2012). The first scale shortcoming is that the use of the term pronunciation is not clearly defined (segmentals only or segmentals combined with suprasegmentals), which causes problems because some errors are more detrimental to comprehensibility than others (Munro & Derwing, 2006). Second, the scales conflate the constructs of comprehensibility and intelligibility, which, as discussed in the introduction of this dissertation, do not have the same meaning. Intelligibility is primarily an issue of decoding L2 speech and is often measured through transcriptions of speech excerpts. Comprehensibility, on the other hand, is related to the degree of difficulty a listener has in understanding a given speaker and is often measured through perception scales, such the ones employed in this study. Third, descriptors are vague and do not offer a clear explanation as to which errors lead to listener difficulty. Finally, scales conflate comprehensibility and accentedness; however, extensive research has revealed that accentedness often does not affect comprehensibility (Derwing & Munro, 1997; Munro & Derwing, 1995).
Given the shortcomings of the current scales used to assess oral proficiency and pronunciation, it is crucial that researchers develop a more reliable scale to measure speaker comprehensibility (Isaacs & Trofimovich, 2012). The language-specific scale used in this study was employed with the intent to shed light on which language features native listeners attend to when judging the comprehensibility of non-native speakers. The findings in this study may contribute to the future development of a more reliable scale, although this is not the purpose of the present study. Chapters 4 (results and discussion) and 5 (conclusion) of this dissertation provide a more detailed discussion of this issue.

**Verbal Protocols**

Verbal protocols were employed while the raters rated the lectures (Appendix E). The purpose of verbal or think-aloud protocols is to explore participants’ cognitive processes (Salinger, Plonka, & Prechelt, 2008) while completing a task. According to Green (1998), “Verbal protocol analysis (VPA) distinguishes itself from other techniques that employ verbal data because, in the case of VPA, inferences are actually made about the cognitive processes that produced the verbalisation” (p. 1). The objective in employing this technique during the ratings is to try to understand the decisions that trained raters make while rating the comprehensibility of the speakers, especially in relation to the features targeted by the Supra Tutor (word stress, rhythm, and intonation).

**ITAs’ Evaluation of the Supra Tutor Questionnaire**

This study also included an evaluation questionnaire (Appendix F) in which ITAs were asked to assess their performance on the Supra Tutor; they were also requested to evaluate the online tutor in terms of usefulness, level of interest, and overall quality of the materials. The questionnaire also prompted participants to indicate whether after the online training they would
prefer to take a pronunciation course face-to-face, completely online, or a hybrid version (half online and half in a classroom). The questionnaire included both Likert-type questions and open-ended questions. Figure 15 shows a screenshot of the questionnaire. The information gathered through the questionnaire was analyzed to evaluate whether ITAs’ perceptions of and interaction with the Supra Tutor had an impact on their comprehensibility improvement.

![Students' evaluation of online materials](image)

**Figure 15.** Screenshot of the ITAs’ evaluation of the Supra Tutor questionnaire

**Procedures**

This study took place in five different stages. Table 6 below provides a summary of the data collection procedures. After the ITAs were recruited, they were given an explanation of the study and were asked to sign the consent form.
In the first stage, a date was set for the pre-test. In order to avoid anxiety as a possible variable, participants were asked to suggest topics that they would be comfortable teaching and that would be easy for undergraduate students from a variety of majors to understand. After topics were suggested, the researcher selected the materials for the participants. Table 4 shows a list of topics for the pre-training and post-training lectures for each participant. Each ITA received the topic for his/her lecture face to face two days prior to the test and was given instructions regarding the format/delivery of the task. They were given help on how to prepare for their lectures and suggestions on what to focus on during the presentations so that anxiety did not become a factor affecting comprehensibility. The ITAs received a brief training on public speaking and classroom management skills to help them select and address main points included in the lecture material. By knowing which points to focus on, they were able to better manage their time during the lecture and focus on more relevant content. They also received a handout (Appendix G) containing before-lecture procedures, during-lecture procedures, and after-lecture procedures. The after-lecture procedures contained an overview of Stage 2 (intervention) and of Stage 3 (post-training lecture). By following these procedures, the researcher encouraged ITAs to feel as comfortable as possible performing the task to minimize the likelihood of undesirable variables (e.g., anxiety) affecting their oral performance. On the day of the scheduled presentation, each ITA delivered his/her seven-minute lecture in a real classroom. Each lecture was videotaped for later analysis.

In the second stage (intervention), ITAs received face-to-face training on how to navigate the online tutor and instructions regarding completion of the modules. They were also given directions on how to fill out the evaluation of the online tutor questionnaire, which was included at the end of the last (review) module. ITAs were also told that they were allowed to go back to
any given section within the four-week time frame to go over the materials again, except for the diagnostic quizzes, which were set to one attempt only. After the data were collected and quiz scores transferred to a hard drive, the quizzes were set to three more attempts so that the ITAs could practice the quizzes for their own benefit. Actually, the entire course was available to them for two additional months after data collection to reward them for their participation in the study.

As an additional incentive, ITAs who participated in all three stages of the study (brief lecture, the online training, and another brief lecture after the training) were eligible to win 1 of 8 $25 gift cards to a store of their choice. In order to partake in the raffle, the ITAs were required to participate in every stage but did not have to complete every task; for instance, they could skip any activity/module in the online training.

Stage three (post-test) took place in the week after the ITAs completed the online training. The same procedures outlined in Stage 1 were followed. This time, the ITAs received a different but related topic. As a matter of fact, several of the lectures were a continuation of the previous lecture, as seen in Table 3.

The next stage of the study involved the rating of the lectures by naïve raters. The data collection from the 178 raters took place in 13 sections of English 150 over a five-day period. Each data collection session lasted for 50 minutes and took place in a computer lab so that the participants could use individual machines for watching the lectures and rating the speakers. PowerPoint slides outlining each step of the process were presented to the participants throughout the data collection process and questions were answered. First, the raters were given an explanation about the study, signed the consent form, and received an ID number for confidentiality. Next, they were instructed to log in to Moodle and complete the background questionnaire. Then, they were given headphones and asked to verify that they were functioning
properly and adjust the volume to their preference. A sound track was available in Moodle for that purpose. After that, a calibration was conducted so that the raters understood the task and the scale to be employed. Part of a lecture on Molecular Biology was played to each class as a whole and the scale was explained based on that sample lecture. Once the calibration was done and questions addressed, the raters were asked to start the rating procedures by watching each of the six lectures only once and rating it immediately after. To avoid sequencing effects, each group rated six videotaped lectures (three pre and three post-training) for overall comprehensibility. The lectures were presented in an alternating order, and only three minutes of each lecture were played to the naïve raters because of time constraints. The original seven-minute lectures had been edited using Camtasia Studio 8 to a length of around three minutes each. The researcher and the teacher for each English 150 section helped with technology issues.

Finally, in Stage 5 the six trained raters rated the speakers. As with the naïve raters, the lectures were presented to the trained raters in alternating order. To avoid sequencing effects, none of the raters rated both pre and post-training lectures from the same speaker. The trained raters were divided into two groups, and each group rated all 12 speakers (six pre and six post-training lectures) for comprehensibility using the scale that includes overall comprehensibility and specific linguistic features. The rating process was conducted with each rater individually over a period of three to four hours on different days. To make the rating more effective and more accurate, the trained raters received training on how to use the scale, and a rater calibration was conducted using one or two videotaped performances from three ITAs who participated in Stage 1 but were unable to complete Stages 2 and 3. Each rater rated 12 lectures individually and verbal protocols (Appendices E and H) were employed. Given that it was unfeasible for the raters to watch each lecture and conduct verbal protocols at the same time, they were instructed
to take notes while watching the lecture. Right after watching each lecture in its entirety, the raters used their notes to rate the speakers on the scale.

The verbal protocols were recorded using a 24 bit Wave MP3 recorder (Edirol R-09). Audacity was used for back-up recording. Rarely did a rater have to re-watch part of a video to complete the rating or the verbal protocols. The purpose of think-alouds was to understand the decisions that trained raters made while rating the comprehensibility of the speakers. More specifically, the purpose to employ verbal protocols was two-fold: 1) to investigate whether the raters commented on the features that were targeted by the Supra Tutor and whether these features were connected to comprehensibility improvement and 2) to analyze which language features the trained raters focused on when rating the speech of the ITAs.
Table 6. Summary of the Data Collection Procedures

<table>
<thead>
<tr>
<th>Stages</th>
<th>Procedures</th>
</tr>
</thead>
</table>
| Stage 1: Pre-test | • ITAs were given the lecture topic two days in advance to prepare their presentation  
                    • ITAs were given instructions regarding the format/delivery of the task  
                    • ITAs were given help on how to prepare for their lectures and what to focus on  
                    • ITAs delivered their lectures, which were videotaped for analysis |
| Stage 2: Intervention | • ITAs received face-to-face training on how to navigate the online tutor and instructions regarding completion of the modules  
                        • ITAs were given four weeks to complete the online training and were told that they were allowed to go back to any given section to go over the materials again, except for the diagnostic quizzes, which were set to one attempt only |
| Stage 3: Post-test | • ITAs were given a related lecture topic two days in advance to prepare their presentation  
                         • ITAs were given instructions regarding the format/delivery of the task  
                         • ITAs were given help on how to prepare for their lectures and what to focus on  
                         • ITAs delivered their lectures, which were videotaped for analysis |
| Stage 4: Ratings 1 | • 13 sections of naïve raters rated six lectures (3 pre and 3 post-training lectures) for overall comprehensibility |
| Stage 5: Ratings 2 | • Two groups of three trained raters rated 12 lectures (6 pre and 6 post-training lectures) for overall comprehensibility and for specific linguistic features  
                            • Each trained rater rated the lectures individually and verbal protocols were employed |

Data Analysis

Both quantitative and qualitative data sources were used. The quantitative data were analyzed using the statistics software package SPSS 20 for Macs. Descriptive and inferential statistics tests were conducted. Before inferential statistical procedures were conducted, the data
derived from the 178 naïve raters were tested for normality to inform the decision as to which kind of statistical procedure(s) (parametric or non-parametric) should be conducted. In order to run parametric tests, independent samples t-tests and paired-samples t-tests in the case of this study, the data do not have to be perfectly distributed. They need to be approximately normally distributed. The data were tested for normality visually through histograms, normal Q-Q (quantile-quantile) plots and box plots, through a calculation of skewness and kurtosis z-values, and through a Shapiro-Wilk test (Henderson, 2006). The data were not perfectly normally distributed but were approximately normal. There were a few outliers in the data; however, these outliers were not removed from the data because they represent valid observations (i.e., individual perception of speaker comprehensibility levels).

As for the qualitative component of this study, data were coded and analyzed using Hyper Research 3.7.1, a qualitative analysis software application. Pre-defined categories and codes were created based on the language-specific comprehensibility scale used by the trained raters. Focused coding was then used to look for data that fit under the themes of interest to develop relevant sub-categories (Charmaz, 2006). Both the researcher and a data analyst who is familiar with the issues under investigation in this study analyzed the qualitative data. She is a language teacher and holds a Ph.D in Applied Linguistics and Technology.

The coding of the qualitative data was done in five stages. The researcher kept a journal of the procedures, discussions, and conclusions to ensure reliability of and an accurate description of the coding procedures. In the first stage, a list of pre-defined categories and codes was created for both qualitative data sets: ITAs’ perceptions of the Supra Tutor (Research Question 2) and verbal protocols from the trained raters (Research Question 3). The pre-defined codes for the ITAs’ perceptions of the online tutor were created based on the open-ended
questions included in the evaluation of the tutor questionnaire as well as on the answers that
ITAs provided to those questions. The list of pre-defined codes for the verbal protocols was
developed based on the 12 items included in the language-specific scale used by the trained
raters.

Second, after the initial codes were developed, I met with the data analyst for about 80
minutes through Skype to explain the research questions, the pre-defined codes and definitions,
and to discuss the coding procedures to follow. We agreed that she would use Microsoft Word
(she had no access to Hyper Research) to perform a sample coding and scheduled another online
meeting to discuss our coding of that sample.

In the third stage, we met online for a second time. Before this second meeting, I sent the
data analyst a document containing an explanation of verbal protocol and the rationale for its
employment in this study (i.e., to provide data that shed light on decisions that raters made when
rating the comprehensibility of L2 speakers), the language-specific comprehensibility scale, the
open-ended questions included in the ITA questionnaire, the codes, the files to be coded, and a
sample of what my coding looked like in Hyper Research. During our second meeting, which
lasted for about 90 minutes, we discussed our sample coding and agreed on changes that should
be made to the initial codes. For instance, we had to include a few codes for themes that emerged
during our sample coding. Tables 7 and 8 display the final list of categories and codes.
Table 7. List of Categories and Codes for Research Question 2

<table>
<thead>
<tr>
<th>Code category</th>
<th>Codes in each category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favorite activity</td>
<td></td>
</tr>
<tr>
<td>Quizzes</td>
<td></td>
</tr>
<tr>
<td>Exercises</td>
<td></td>
</tr>
<tr>
<td>Exercises with <em>Audacity</em> or <em>Praat</em></td>
<td></td>
</tr>
<tr>
<td>Lectures</td>
<td></td>
</tr>
<tr>
<td>Least favorite activity</td>
<td></td>
</tr>
<tr>
<td>Quizzes</td>
<td></td>
</tr>
<tr>
<td>Exercises</td>
<td></td>
</tr>
<tr>
<td>Lectures</td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>Favorite aspects of the tutor</td>
<td></td>
</tr>
<tr>
<td>Non-inhibiting environment</td>
<td></td>
</tr>
<tr>
<td>Flexibility (time)</td>
<td></td>
</tr>
<tr>
<td>Self-pace</td>
<td></td>
</tr>
<tr>
<td>Variety of materials and activities</td>
<td></td>
</tr>
<tr>
<td>Learning new technology</td>
<td></td>
</tr>
<tr>
<td>Effectiveness of materials</td>
<td></td>
</tr>
<tr>
<td>No grade pressure</td>
<td></td>
</tr>
<tr>
<td>Instant feedback on perception exercises</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>Least favorite aspects of the tutor</td>
<td></td>
</tr>
<tr>
<td>Technology</td>
<td></td>
</tr>
<tr>
<td>Exercises</td>
<td></td>
</tr>
<tr>
<td>Software (<em>Audacity</em> or <em>Praat</em>)</td>
<td></td>
</tr>
<tr>
<td>Feedback</td>
<td></td>
</tr>
<tr>
<td>Length of training</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>Challenges of online classes for international students</td>
<td></td>
</tr>
<tr>
<td>Lack of interaction with instructor</td>
<td></td>
</tr>
<tr>
<td>Interaction with technology</td>
<td></td>
</tr>
<tr>
<td>Material organization</td>
<td></td>
</tr>
<tr>
<td>Motivation</td>
<td></td>
</tr>
<tr>
<td>Feedback</td>
<td></td>
</tr>
<tr>
<td>Transfer to new contexts</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>Suggestions for change</td>
<td></td>
</tr>
<tr>
<td>Chat room</td>
<td></td>
</tr>
<tr>
<td>Shorter lectures</td>
<td></td>
</tr>
<tr>
<td>Feedback to each other</td>
<td></td>
</tr>
<tr>
<td>Highlight most important activities</td>
<td></td>
</tr>
<tr>
<td>Add lip movement</td>
<td></td>
</tr>
<tr>
<td>More examples</td>
<td></td>
</tr>
<tr>
<td>More production exercises</td>
<td></td>
</tr>
</tbody>
</table>
Table 7 continued

<table>
<thead>
<tr>
<th>Type of class preferred</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fully online</td>
</tr>
<tr>
<td></td>
<td>Face-to-face</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Impressions of the tutor</th>
<th>Usefulness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Level of interest</td>
</tr>
<tr>
<td></td>
<td>Quality of materials</td>
</tr>
<tr>
<td></td>
<td>Other</td>
</tr>
</tbody>
</table>

Table 8. List of Categories and Codes for Research Question 3

<table>
<thead>
<tr>
<th>Code category</th>
<th>Codes in each category (based on scale; 1=positive, 9=negative ratings)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consonants</td>
<td>Consonants 1-3</td>
</tr>
<tr>
<td></td>
<td>Consonants 4-6</td>
</tr>
<tr>
<td></td>
<td>Consonants 7-9</td>
</tr>
<tr>
<td></td>
<td>Consonant insertion (e.g., Campu’s)</td>
</tr>
<tr>
<td></td>
<td>Final consonant omission</td>
</tr>
<tr>
<td>Vowels</td>
<td>Vowels 1-3</td>
</tr>
<tr>
<td></td>
<td>Vowels 4-6</td>
</tr>
<tr>
<td></td>
<td>Vowels 7-9</td>
</tr>
<tr>
<td></td>
<td>Vowel insertion (e.g., Istuff)</td>
</tr>
<tr>
<td></td>
<td>Vowel omission</td>
</tr>
<tr>
<td>Word stress</td>
<td>Word stress 1-3</td>
</tr>
<tr>
<td></td>
<td>Word stress 4-6</td>
</tr>
<tr>
<td></td>
<td>Word stress 7-9</td>
</tr>
<tr>
<td>Stressed syllables/words in a sentence</td>
<td>Stressed words 1-3</td>
</tr>
<tr>
<td></td>
<td>Stressed words 4-6</td>
</tr>
<tr>
<td></td>
<td>Stressed words 7-9</td>
</tr>
<tr>
<td>Unstressed syllables/words in a sentence</td>
<td>Unstressed words 1-3</td>
</tr>
<tr>
<td></td>
<td>Unstressed words 4-6</td>
</tr>
<tr>
<td></td>
<td>Unstressed words 7-9</td>
</tr>
<tr>
<td></td>
<td>Syllable omission</td>
</tr>
<tr>
<td>Final intonation</td>
<td>Final intonation 1-3</td>
</tr>
<tr>
<td></td>
<td>Final intonation 4-6</td>
</tr>
<tr>
<td></td>
<td>Final intonation 7-9</td>
</tr>
<tr>
<td>Voice range</td>
<td>Voice range 1-3</td>
</tr>
<tr>
<td></td>
<td>Voice range 4-6</td>
</tr>
<tr>
<td></td>
<td>Voice range 7-9</td>
</tr>
<tr>
<td>Unnatural pauses/hesitations</td>
<td>Unnatural pauses/hesitations 1-3</td>
</tr>
<tr>
<td></td>
<td>Unnatural pauses/hesitations 4-6</td>
</tr>
<tr>
<td></td>
<td>Unnatural pauses/hesitations 7-9</td>
</tr>
<tr>
<td>Grammar use</td>
<td>Grammar 1-3</td>
</tr>
</tbody>
</table>
Next, once the categories and codes were fine-tuned, we coded twenty percent of the data individually and inter-rater reliability was calculated. A simple percentage agreement calculation was used. Table 9 shows the inter-rater reliability percentage for each data set. For the ITAs’ perceptions of the Supra Tutor data, we had a total of 147 threads (147 codes) applied. Out of those 147 instances, we had 17 disagreements. For this data set, rather than discrepancy regarding the application of one code over another, our disagreements were often the application of a code that one of us did not employ or the addition of a code that also fit under a different category. As for the verbal protocols data set, we coded a total of 169 threads (169 codes applied). We disagreed on 15 instances. Given that the codes for this data set were generated based on the language-specific scale (Table 8), they were somewhat straightforward. However, the category “Listener perception” was the category responsible for our disagreements. That is, I added that category to several of the instances coded by the analyst. For instance, we both coded “Combined with his thought groups, as a student I think I would – his thought groups are extremely broken. I would get really frustrated very fast and then I would just quit trying” (Rater Four, Lecture One, Appendix H) as “Unnatural pauses and hesitations 7-9”. However, I added
“Listener perception” as a second code to this thread because this specific thread relates to how the rater perceived the speech being judged.

Table 9. Inter-rater Reliability for the Data Sets Coded by the Researcher and the Data Analyst

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Threads</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITAs’ perceptions of the Supra Tutor</td>
<td>30</td>
<td>89%</td>
</tr>
<tr>
<td>Trained raters verbal protocols</td>
<td>169</td>
<td>91%</td>
</tr>
</tbody>
</table>

Research Question 1 (What is the impact of the online pronunciation tutor on the comprehensibility of ITAs?) was addressed through the quantitative analysis (descriptive and inferential statistics) of the data derived from the overall comprehensibility ratings from 178 domestic undergraduate students (naïve raters) enrolled in 13 sections of English 150, Critical Thinking and Communication, at Iowa State University. Data gathered from the Supra Tutor activity reports were used to support the findings yielded by the quantitative data.

The quantitative and qualitative data for the analysis of Research Question 2 (How do the ITAs evaluate the online pronunciation tutor in regards to usefulness, level of interest, and quality of the materials?) were derived from three different sources: a numeric and open-ended online questionnaire that the ITAs completed at the end of the training, emails from the participants, and a forum activity included in the online tutor. The forum activity prompted participants to write about their experience as graduate students in America and about their experience participating in the online pronunciation training. The qualitative data were coded (Table 7) by the researcher and by a data analyst as described above.

Both quantitative and qualitative data were used to answer Research Question 3 (What is the contribution of suprasegmental features (i.e., ratings of word stress, rhythm, and intonation) to comprehensibility ratings?). The data for the analysis of this research question were derived from the ratings by six trained raters (quantitative) and from verbal protocols (e.g., Zielinski,
that were audio-recorded while the raters performed their individual rating of each speaker (qualitative). The quantitative data were analyzed through descriptive statistics alone since the number of rater in each group was too small (N = 3) for inferential statistics. The qualitative data (verbal protocols) were transcribed in their entirety (a total of 72 Microsoft Word files), coded (Table 8) by the researcher and the data analyst as discussed above, and then analyzed using Hyper Research. Table 10 provides a summary of the research questions and data analysis instruments.

**Table 10. Summary of Research Questions and Data Collection Instruments**

<table>
<thead>
<tr>
<th>Research question</th>
<th>Data source</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What is the impact of the online pronunciation tutor on the comprehensibility of ITAs?</td>
<td><strong>QUAN</strong> Pre and post-test comprehensibility ratings from 178 naïve raters</td>
<td><strong>QUAN</strong> Descriptive statistics Paired-samples <em>t</em>-tests Independent samples <em>t</em>-tests</td>
</tr>
<tr>
<td></td>
<td><strong>qual</strong> Supra Tutor activity reports</td>
<td><strong>qual</strong> Holistic account</td>
</tr>
<tr>
<td>2. How do the ITAs evaluate the online pronunciation tutor in regards to usefulness, level of interest, and quality of the materials?</td>
<td><strong>QUAN</strong> Quantitative portion of the ITAs’ evaluation questionnaire</td>
<td><strong>QUAN</strong> Descriptive statistics</td>
</tr>
<tr>
<td></td>
<td><strong>qual</strong> Open-ended questions in the ITAs’ evaluation questionnaire Follow-up email responses Forum activity</td>
<td><strong>qual</strong> Pre-defined coding Focused coding</td>
</tr>
<tr>
<td>3. What is the contribution of suprasegmental features (i.e., ratings of word stress, rhythm, and intonation) to comprehensibility ratings?</td>
<td><strong>QUAN</strong> Pre and post-test ratings from six trained raters</td>
<td><strong>QUAN</strong> Descriptive statistics</td>
</tr>
<tr>
<td></td>
<td><strong>qual</strong> Think-aloud protocol data from trained raters</td>
<td><strong>qual</strong> Pre-defined coding Focused coding</td>
</tr>
</tbody>
</table>
Chapter Summary

This chapter described the data collection methods and materials employed in this dissertation. It provided a detailed description of the research design, setting, participants, and the Supra Tutor, the online pronunciation course developed for this study. This chapter also provided a detailed explanation of the materials and procedures used in the data collection process. An explanation of the research questions and data analysis method for each question concluded the chapter. The next section of this dissertation, Chapter 4 presents and discusses the findings of this research.
CHAPTER 4: RESULTS AND DISCUSSION

This chapter presents and discusses the findings for each research question. First, the chapter discusses the impact of online pronunciation instruction on the comprehensibility of ITAs. Findings indicate that the Supra Tutor was effective in providing training in the recognition and proper application of suprasegmental features in the English Language. Next, the chapter presents the findings regarding ITAs’ perceptions and evaluations of the Supra Tutor. The results show that the ITAs found the materials to be useful, informative, engaging, motivating, and of high quality. After that, the chapter focuses on trained raters’ evaluation of the comprehensibility of the speakers and on how the use of suprasegmentals impacted their perceptions of ITAs’ comprehensibility. Results indicated that raters attend to suprasegmentals, especially word stress. However, there was no clear connection between suprasegmental ratings and comprehensibility ratings. Finally, the chapter concludes with a discussion of other language features that the trained raters attended to when judging the ITAs’ speech. It was found that raters focused on various language features (e.g., grammar, fluency) in addition to suprasegmentals. However, some segmentals, vowels in particular, discriminated among levels of proficiency.

Research Question 1: Impact of the Online Pronunciation Tutor on ITA Comprehensibility

This question investigated whether naïve listeners perceived changes in the comprehensibility of 12 ITAs as a result of their focused self-study using the Supra Tutor, a four-week online pronunciation tutor focusing on word stress, rhythm, and intonation.

This question was answered through ratings collected from the naïve raters and through a paired-samples $t$-test and independent samples $t$-tests. In addition, data from the Supra Tutor activity reports for each ITA were used to support the findings from the quantitative data. Table
11 summarizes the groups and the speakers that each group rated. As seen in the table, none of the groups rated pre and post lectures of the same speaker. The purpose of this data collection design was to avoid familiarity effects. In order to interpret the results, it is essential to take the scale used into account. The holistic scale ranged from 1 to 9, with lower numbers representing positive ratings (e.g., extremely easy to understand), and higher numbers representing negative ratings (e.g., impossible to understand). Hence, lower scores represent more positive evaluations of the speakers’ comprehensibility, and higher scores represent more negative judgments of comprehensibility.

**Table 11. Data Collection Design for the Naïve Raters**

<table>
<thead>
<tr>
<th>Groups of naïve raters</th>
<th>N</th>
<th>Lectures rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>42</td>
<td>Speaker 1 pre</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 2 post</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 3 pre</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 4 post</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 5 pre</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 6 post</td>
</tr>
<tr>
<td>Group 2</td>
<td>38</td>
<td>Speaker 7 pre</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 8 post</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 9 pre</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 10 post</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 11 pre</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 12 post</td>
</tr>
<tr>
<td>Group 3</td>
<td>42</td>
<td>Speaker 1 post</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 2 pre</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 3 post</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 4 pre</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 5 post</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 6 pre</td>
</tr>
<tr>
<td>Group 4</td>
<td>56</td>
<td>Speaker 7 post</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 8 pre</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 9 post</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 10 pre</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 11 post</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaker 12 pre</td>
</tr>
</tbody>
</table>
Table 12 shows the means and standard deviations for pre and post-treatment comprehensibility ratings assigned by each group of naïve raters. Based on descriptive statistics alone, Speakers 2 (L1 Amharic), 3 (L1 Farsi), 5 (L1 Spanish), 8 (L1 Arabic), and 9 (L1 Portuguese) received more positive comprehensibility ratings on their post-test (post-treatment) lectures. On the other hand, Speakers 1 (L1 Farsi), 7 (L1 Hindi), 11 (L1 Chinese), and 12 (L1 Chinese) received roughly the same comprehensibility ratings on both pre and post-tests. Interestingly, Speakers 4 (L1 Farsi), 6 (L1 Chinese), and 10 (L1 Chinese) received more negative comprehensibility ratings on their post-test lectures. Before independent samples t-tests (Table 13) were conducted to examine the significance of the outcomes, a paired-samples t-test was run to compare group means (pre and post-test ratings) to examine how the 12 speakers performed as a group. The paired-samples t-test results ($t(11) = .919, p = .378$) indicated that there was no significant comprehensibility improvement when the ratings were analyzed for the speakers as a group (pre $M = 4.54$ ($SD = 1.26$), post $M = 4.32$ ($SD = 1.38$), $p = .378$). Following is a discussion of the comprehensibility analysis for each of the 12 speakers.

Table 12. Descriptive Statistics of Pre and Post-test Comprehensibility Ratings by the Naïve Raters

<table>
<thead>
<tr>
<th>Speaker (S)</th>
<th>L1</th>
<th>Pre-test</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$N$</td>
<td>Mean</td>
</tr>
<tr>
<td>S1</td>
<td>Farsi</td>
<td>42</td>
<td>4.00</td>
</tr>
<tr>
<td>S2</td>
<td>Amharic</td>
<td>42</td>
<td>6.45</td>
</tr>
<tr>
<td>S3</td>
<td>Farsi</td>
<td>42</td>
<td>3.24</td>
</tr>
<tr>
<td>S4</td>
<td>Farsi</td>
<td>42</td>
<td>3.90</td>
</tr>
<tr>
<td>S5</td>
<td>Spanish</td>
<td>42</td>
<td>2.00</td>
</tr>
<tr>
<td>S6</td>
<td>Chinese</td>
<td>42</td>
<td>6.07</td>
</tr>
<tr>
<td>S7</td>
<td>Hindi</td>
<td>38</td>
<td>4.58</td>
</tr>
<tr>
<td>S8</td>
<td>Arabic</td>
<td>56</td>
<td>5.38</td>
</tr>
<tr>
<td>S9</td>
<td>Portuguese</td>
<td>38</td>
<td>4.32</td>
</tr>
<tr>
<td>S10</td>
<td>Chinese</td>
<td>56</td>
<td>5.63</td>
</tr>
<tr>
<td>S11</td>
<td>Chinese</td>
<td>38</td>
<td>3.90</td>
</tr>
<tr>
<td>S12</td>
<td>Chinese</td>
<td>56</td>
<td>5.00</td>
</tr>
</tbody>
</table>
Independent samples \( t \)-tests (Table 13) assessed whether comprehensibility means differed significantly for the speakers regarding the ratings they received on their pre and post-tests. As mentioned in Chapter 3, preliminary data screening indicated that the ratings assigned by the naïve raters were not perfectly normally distributed. However, based on normality tests (histograms, Q-Q plots, box plots, skewness and kurtosis z-value calculations, and Shapiro-Wilk tests) the departure from normality was not judged serious enough to require the use of a nonparametric test. The assumption of homogeneity of variance was assessed by the Levene’s \( F \) test, which indicated no significant violation of the equal variance assumption for Speaker 1 \( F(82) = .353, p = .554 \); Speaker 3 \( F(82) = 2.08, p = .152 \); Speaker 5 \( F(82) = .433, p = .512 \); Speaker 6 \( F(82) = .806, p = .372 \); Speaker 7 \( F(92) = .130, p = .719 \); Speaker 8 \( F(92) = .174, p = .678 \); Speaker 9 \( F(92) = 2.93, p = .090 \); Speaker 10 \( F(92) = .919, p = .340 \); Speaker 11 \( F(92) = .585, p = .446 \); and Speaker 12 \( F(92) = .060, p = .807 \). However, for Speaker 2 and for Speaker 4 equal variances were not assumed \( (F(82) = 6.69, p = .011 \) and \( F(82) = 5.04, p = .027 \), respectively).
Table 13. Independent Samples T-tests for Mean Comparison Between Pre and Post-tests

<table>
<thead>
<tr>
<th>Speaker (S)</th>
<th>Test</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>t-test (Equal variances assumed, except for S2 and S4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Pre</td>
<td>42</td>
<td>4.00</td>
<td>1.65</td>
<td>-1.222</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>42</td>
<td>4.38</td>
<td>1.56</td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>Pre</td>
<td>42</td>
<td>6.45</td>
<td>1.46</td>
<td>-3.587</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>42</td>
<td>5.14</td>
<td>1.85</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>Pre</td>
<td>42</td>
<td>3.24</td>
<td>1.28</td>
<td>2.231</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>42</td>
<td>2.67</td>
<td>1.05</td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>Pre</td>
<td>42</td>
<td>3.90</td>
<td>1.41</td>
<td>4.177</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>42</td>
<td>4.86</td>
<td>1.80</td>
<td></td>
</tr>
<tr>
<td>S5</td>
<td>Pre</td>
<td>42</td>
<td>2.00</td>
<td>1.03</td>
<td>1.121</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>42</td>
<td>1.76</td>
<td>.90</td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>Pre</td>
<td>42</td>
<td>6.07</td>
<td>1.71</td>
<td>1.635</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>42</td>
<td>6.60</td>
<td>1.47</td>
<td></td>
</tr>
<tr>
<td>S7</td>
<td>Pre</td>
<td>38</td>
<td>4.58</td>
<td>1.61</td>
<td>.284</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>56</td>
<td>4.48</td>
<td>1.60</td>
<td></td>
</tr>
<tr>
<td>S8</td>
<td>Pre</td>
<td>56</td>
<td>5.38</td>
<td>1.69</td>
<td>-4.650</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>38</td>
<td>3.74</td>
<td>1.65</td>
<td></td>
</tr>
<tr>
<td>S9</td>
<td>Pre</td>
<td>38</td>
<td>4.32</td>
<td>2.05</td>
<td>2.921</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>56</td>
<td>3.16</td>
<td>1.75</td>
<td></td>
</tr>
<tr>
<td>S10</td>
<td>Pre</td>
<td>56</td>
<td>5.63</td>
<td>1.75</td>
<td>1.243</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>38</td>
<td>6.11</td>
<td>1.95</td>
<td></td>
</tr>
<tr>
<td>S11</td>
<td>Pre</td>
<td>38</td>
<td>3.90</td>
<td>1.81</td>
<td>-0.339</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>56</td>
<td>3.95</td>
<td>1.84</td>
<td></td>
</tr>
<tr>
<td>S12</td>
<td>Pre</td>
<td>56</td>
<td>5.00</td>
<td>1.91</td>
<td>.213</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>38</td>
<td>5.05</td>
<td>2.05</td>
<td></td>
</tr>
</tbody>
</table>

*Note. 95% CI of the difference (*Significant at p<.05)

As seen in Table 13, Speakers 2, 3, 8, and 9 showed significant improvement in their comprehensibility ratings while Speaker 4’s comprehensibility was judged to have become significantly worse on her post-test (pre $M = 3.90$ ($SD = 1.41$), post $M = 4.86$ ($SD = 1.80$), $p = .000$). The comprehensibility of all the other speakers (1, 5, 6, 7, 10, 11, and 12) was judged to have remained the same after the online training.

Given that RQ 1 included only holistic quantitative data, it was not possible to determine the reasons for the independent samples $t$-test outcomes. However, the activity reports generated
by the Supra Tutor may provide some explanation as to why some speakers showed comprehensibility improvement whereas others did not and why Speaker 4 received more negative comprehensibility ratings on her post-test lecture. The next section in this chapter will investigate the ITAs’ perception of the tutor and will establish connections between the comprehensibility findings for RQ 1 and ITAs’ interaction with and evaluation of the Supra Tutor. However, following is a brief analysis and discussion of the activity reports for the ITAs (speakers) based on the three independent t-tests outcomes: speakers who improved their comprehensibility (2, 3, 8 and 9), the speaker who received significantly more negative post-test comprehensibility ratings (Speaker 4), and the speakers whose comprehensibility remained the same after the online training (1, 5, 6, 7, 10, 11, and 12).

Before discussing the activity reports (generated by Moodle), a few points need to be made. First, the Supra Tutor focused on word stress, rhythm, and intonation. It did not include segmentals (vowels and consonants), grammar, or vocabulary. While the ITAs may have improved their suprasegmentals, they were rated for overall comprehensibility. That is, other language features that were not included in the Supra Tutor may have played a role in how comprehensible the naïve raters found the speakers to be either before or after the training. This possibility is impossible to rule out in this kind of production study. That is one factor that motivated Research Question 3, which examined the language features that trained raters attended to when judging comprehensibility. Second, while the activity reports displayed numbers of views, scores, and day and time of access, they do not provide an account of time or quality of time spent on each task. Therefore, time on task or full completion of a given task, which may have had an impact on how well the ITAs did, is not accounted for in the following discussion.
Analysis of the Supra Tutor Activity Reports

Speakers who improved their comprehensibility

As revealed by the independent samples t-tests (Table 13), speakers 2, 3, 8, and 9 improved their comprehensibility significantly. An analysis of their Supra Tutor activity reports showed that speakers 2, 8, and 9 accessed all the instructional materials multiple times and performed and submitted all perception and production exercises from all four modules. Speaker 3’s use of the tutor showed active use for the first three modules (word stress, rhythm, and intonation) and submission of all activities in those three modules. However, Speaker 3 did not complete the review module in its entirety by the deadline (end of data collection period). He finished the general review quiz and the word stress review quiz but not the remaining activities in the module, which included review perception and production tasks. However, as in the case of Speakers 2, 8, and 9, Speaker 3 performed and submitted all production exercises, which made use of either Audacity or Praat or a combination of both.

The activity reports also showed that these speakers obtained relatively low diagnostic quiz scores, especially in regards to rhythm (average = 59%). However, they received a higher score on the rhythm review quiz (76%). Although Speaker 3 did not complete the rhythm review quiz, he completed the general review quiz (word stress, rhythm, and intonation combined), obtaining a score of 28.5 out of 35 possible points (81.4%). An improvement in the perception review quizzes indicates that the speakers seemed to have noticed their shortcomings and decided to revisit the instructional materials and properly follow the instructions during the perception and production exercises. It could be argued that they increased their awareness of English suprasegmentals, which was one of the primary goals of the Supra Tutor.

Speaker who received more negative post-test comprehensibility ratings
Speaker 4 received statistically significant ($t(82) = 4.177, p = .001$) lower comprehensibility ratings on her post-test (pre $M = 3.90$ ($SD = 1.41$), post $M = 4.86$ ($SD = 1.80$)). An analysis of her Supra Tutor activity report showed that she viewed most of the instructional materials once or twice and completed the diagnostic quizzes and perception exercises. She failed to submit several of the production activities, which were designed to practice the knowledge acquired through instructional lectures and perception tasks. For instance, Speaker 4 did not submit either of the word stress production tasks nor did she submit two of the intonation production exercises. In addition, instead of using Audacity and Praat for some production exercises that required the use of those two platforms for aural and visual feedback, she submitted audio files recorded on her cell phone. If learners, ITAs in this case, do not perform the production tasks it is unlikely that they will be able to use pronunciation features successfully. A more detailed discussion regarding Speaker 4 and the other speakers is provided in the next two sections (Research Question 2 and Research Question 3).

*Speakers whose comprehensibility remained the same after training*

When it comes to the speakers who did not show comprehensibility improvement (Speakers 1, 5, 6, 7, 10, 11, and 12), the Supra Tutor activity reports revealed that although the majority of these speakers completed most of the activities in the tutor, some (e.g., Speakers 1 and 10) completed several of the activities in each module on the same day, generally in three-hour blocks on the weekend before each module completion was due (participants were assigned a module per week, four modules over four weeks). When it comes to pronunciation skills, it may be best for learners to practice for a short amount of time each day rather than for a longer period of time during one or two days per week (Derwing, 2010).
Second, the word stress, rhythm, and intonation video-lectures were designed to raise speakers’ awareness of English suprasegmental features in terms of structure and discourse functions. Language awareness is one key element to help speakers improve their comprehensibility (Morley, 1991). Some of the speakers who received relatively low scores on diagnostic quizzes, which should be an indication that they needed to focus their attention on activities related to their areas of weakness, did not access instructional materials more than once (e.g., Speakers 1, 6, 10).

Third, although Speaker 5 completed most of the activities as prompted, he did not show comprehensibility improvement. This is most likely due to the fact that his English oral proficiency was already advanced before the online training (pre $M = 2.00$ ($SD = 1.03$), post $M = 1.76$ ($SD = .90$)).

Fourth, Speaker 7 received similar comprehensibility ratings on her pre ($M = 4.58$ ($SD = 1.61$)) and post ($M = 4.48$ ($SD = 1.60$)) tests. However, the activity report showed that she was diligent in using the Supra Tutor. She accessed the instructional materials multiple times and submitted all perception and production exercises. Also, she received relatively low scores on the diagnostic quizzes but showed remarkable improvement in the review quizzes. A possible explanation is that language features and even non-language features (e.g., speed) may have impacted naïve raters’ perception of this speakers’ comprehensibility. Also, Speaker 7 is a native speaker of Hindi, but English is a comfortable, almost native language for her. However, her English variety is closer to British English than to American English. Although her English is fluent, she speaks a variety of English that is hard to change even with diligent practice.

Finally, although some speakers showed active interaction with the Supra Tutor (e.g., Speakers 7 and 11), they did not show comprehensibility improvement. For instance, Speaker 11
completed all the activities in the tutor and performed the perception and production exercises properly following instructions (e.g., using Praat for intonation practice) but received roughly the same ratings on her pre and post-tests (pre $M = 3.90$ ($SD = 1.81$), post $M = 3.95$ ($SD = 1.84$), $p = .735$). A possible explanation might be that the number of perception and production exercises included in the tutor may not have been sufficient to allow her to practice the new patterns that she learned about in the Supra Tutor. Additionally, not every learner will improve quickly. Time may have been a factor considering that the online training was administered over a four-week period.

As mentioned previously, only quantitative data (holistic comprehensibility scale) were collected from the 178 domestic undergraduate students in this study. The next section in this chapter discusses the speakers’ (ITAs) perceptions of and interaction with the Supra Tutor. This chapter concludes with a discussion of the findings for Research Question 3: What is the contribution of suprasegmental features (i.e., ratings of word stress, rhythm, and intonation) to comprehensibility ratings? This research question, which draws on verbal protocol data from trained raters, is meant to shed some light on the findings in this section (RQ 1).

**Research Question 2: ITAs’ Evaluation of the Online Pronunciation Tutor**

How did the ITAs evaluate the online pronunciation tutor in regards to usefulness, level of interest, and quality of the materials? The purpose of this research question was to provide insight into the connection between improvement and motivation, as measured by ITAs’ completions of the online training. It is noteworthy, however, that while Research Question 1 used the Supra Tutor activity reports to support the quantitative findings, Research Question 2 examines ITAs’ perceptions of their performance and their evaluation of the Supra Tutor. That
is, ITAs’ perceptions of their performance may or may not match the findings discussed in the previous section (RQ 1).

The data for this specific question come from three different sources: 1) a numeric and open-ended online questionnaire that the ITAs completed at the end of the training; 2) emails from the participants; and 3) a forum activity included in the online tutor. The forum activity prompted participants to write about their experience as graduate students in the U.S. and about their experience participating in the online training.

As discussed in the previous section, four of the ITAs in this study (Speaker 2, Speaker 3, Speaker 8, and Speaker 9) showed significant improvement in their post-test comprehensibility ratings as revealed by the independent samples \( t \)-tests (\( t(82) = -3.587, p = .000; t(82) = 2.231, p = .028; t(92) = -4.650, p = .001; \) and \( t(92) = 2.921, p = .004 \), respectively). One ITA (Speaker 5) received more positive comprehensibility ratings on his post-test (pre \( M = 2.00 \) (SD = 1.03), post \( M = 1.76 \) (SD = .90)), but that difference was not significant (\( t(82) = 1.121, p = .265 \)). Four of the speakers (1, 7, 11, and 12) received roughly the same comprehensibility ratings on both pre and post-tests. Speakers 4, 6, and 10, on the other hand, received more negative comprehensibility ratings on their post-tests. However, the only significant difference was found for Speaker 4 (pre \( M = 3.90(SD = 1.41), \) post \( M = 4.86(SD = 1.80); t(82) = 4.177, p = .001 \).

An analysis of the Supra Tutor activity report in the previous section (RQ 1) revealed that some ITAs made better use of the Supra Tutor than others. For instance, the ITAs that showed significant improvement accessed instructional materials numerous times and performed perception and production tasks as prompted (e.g., using Audacity and Praat for visual feedback).
Next, I discuss the ITAs’ evaluation of the online tutor. The questionnaire (Appendix C) prompted participants to evaluate their own performance during the online training and to assess the tutor in terms of usefulness, level of interest, and quality and clarity of the materials (Likert-type questions). It also asked them to indicate their favorite activity, least favorite activity, what they liked the most about the Supra Tutor, their least favorite aspects of the tutor, and suggestions for improvement (open-ended questions). Finally, the questionnaire prompted ITAs to indicate, based on their experience with the Supra Tutor, what type of pronunciation course they would likely take in the future: face-to-face, completely online, or a hybrid course (part online and part in a regular classroom) and why. This last part of the questionnaire was tied to participants’ perceptions of the tutor in terms of usefulness, level of interest, and quality of the materials. In other words, how would they compare online pronunciation training using the Supra Tutor to classroom instruction (which they did not receive in this study)?

Using both quantitative and qualitative data, a discussion of the ITAs’ perceptions and evaluations of the Supra Tutor is provided in two different respects: 1) their perceptions of their own performance in the tutor and 2) their evaluation of the Supra Tutor in regards to usefulness, level of interest, and overall quality and clarity of the materials. In order to maintain consistency throughout this chapter, the ITAs are addressed as “speakers” as in the case of Research Question 1.

**Speakers’ perception of their performance and interaction with the Supra Tutor**

Table 14 provides descriptive statistics for the user performance variables included in the speakers’ evaluation of the Supra Tutor questionnaire. In order to maintain consistency and avoid confusion, the scale used in the questionnaire follows the same pattern as the scales employed to assess speaker comprehensibility. That is, lower numbers represent positive evaluations (e.g., 1 =
extremely useful) and higher numbers represent more negative evaluations (e.g., $5 = \text{not at all useful}$). As seen in the table, overall the speakers indicated they thought they were regular (i.e., dedicated) in their use of the Supra Tutor ($M = 2.00 \ (SD = .85)$). They also reported using the tutor to learn about pronunciation, new technologies, and new learning strategies ($M = 1.41 \ (SD = .51)$). In regards to their perception of the effectiveness of the online training over traditional (face-to-face) instruction, the speakers agreed that they learned more about English pronunciation using the tutor than they believe they would have in a regular classroom ($M = 1.54 \ (SD = .52)$).

**Table 14.** Descriptive Statistics for Perceived User Performance in the Supra Tutor

<table>
<thead>
<tr>
<th>Variable</th>
<th>Descriptors</th>
<th>Mean ($N = 12$)</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of dedication</td>
<td>1 = extremely dedicated, 2 = dedicated, 3 = somewhat dedicated, 4 = not very dedicated, 5 = not at all dedicated</td>
<td>2.00</td>
<td>.85</td>
</tr>
<tr>
<td>Used the tutor to learn about pronunciation and new strategies and technology</td>
<td>1 = definitely true, 2 = true, 3 = somewhat true, 4 = not at all true</td>
<td>1.41</td>
<td>.51</td>
</tr>
<tr>
<td>Supra Tutor vs. F2F classes</td>
<td>Through the online materials I learned____ than I would in a regular classroom. 1 = much more, 2 = a little more, 3 = as much as, 4 = less than, 5 = much less</td>
<td>1.54</td>
<td>.52</td>
</tr>
</tbody>
</table>

As seen in Figure 16, most speakers reported having been dedicated during the online training. This question included five options (1 = extremely dedicated, 2 = dedicated, 3 = somewhat dedicated, 4 = not very dedicated, 5 = not at all dedicated). Speakers 2, 7, and 8 reported being extremely dedicated. The connection between self-rated dedication and rated comprehensibility improvement was not consistent. Speakers 2 and 8 showed significant comprehensibility improvement on their post-tests (see Table 13). Speaker 3 claimed to be
somewhat dedicated but also received significantly better comprehensibility ratings on his post-test. Although his Supra Tutor activity report showed that he did not finish the review module, module 4, he was very active in the first three modules and completed all perception and production exercises, properly following directions. For instance, in an email correspondence Speaker 4 informed me that, “I finished module 3. I could not speak louder for the last assignment since I may bother my roommate with my Hollywoodian accent at midnight [smiley face]”. The activity that he was referring to was the last production task in the intonation module. It prompted speakers to watch a scene from “Seinfeld” (George’s handicapped speech) as many times as they wanted and “imitate” the speakers in regards to rhythm, prominence, and intonation. The speakers were to record the dialogue in Audacity and listen to their own recording to monitor their progress.

![ Speakers' perceptions of their performance in the Supra Tutor ]

Figure 16. Speakers' perceptions of their own performance in the Supra Tutor

The next question prompted the speakers to indicate on a scale of 1 to 5 how much advantage they took of the Supra Tutor to learn about pronunciation, new technologies, and new
learning strategies (1 = definitely true, 2 = true, 3 = somewhat true, 4 = not at all true). Six of the speakers (Speakers 2, 4, 6, 7, 8, and 10) reported taking full advantage of the tutor. However, as pointed out in the previous section (RQ1), instead of using Audacity and Praat for production practice, Speaker 4 (the speaker who received significantly lower comprehensibility ratings on her post-test) used her cell phone to record the files. The purpose of those two platforms was to allow speakers to compare output with input. That is, because the Supra Tutor was fully online, production feedback was provided through aural and visual feedback. By not using Praat and Audacity as prompted, Speaker 4 failed to draw the output/input comparison and to properly use self-monitoring strategies. When asked about her least favorite activity in the tutor, Speaker 4 claimed that, “For me because I am usually busy, the parts about installing new software and figure out how they work”. In order to avoid user frustration during the training, very short screen capture tutorials were created to show speakers how to install and use Audacity and Praat for aural and visual feedback. While Speaker 4 accessed those tutorials once or twice, she chose not to use the software as suggested.

The questionnaire also asked the speakers to compare knowledge gain through the tutor as compared to face-to-face pronunciation classes (Through the online materials I learned... 1 = much more, 2 = a little more, 3 = as much as, 4 = less than, 5 = much less than I would in a regular classroom. It is important to note that the speakers did not have a face-to-face pronunciation class. Thus, the information they provided for this question is based on what they believed each type of instruction would offer them.

As seen in Figure 16, all 12 speakers indicated that they either learned much more (N = 5) or a little more than they would in a face-to-face class (N = 7). Next, the speakers were asked to indicate, based on their experience with the Supra Tutor, what types of pronunciation class (fully
online, face-to-face, or hybrid) they would take in the future and why. Table 15 displays the type of instruction each speaker chose and their rationale for their choices. Five speakers chose “fully online classes”, six chose “hybrid”, and one Speaker chose both “fully online” and “hybrid” types of pronunciation instruction. None of the speakers indicated that they would prefer face-to-face instruction, which suggests that they found the Supra Tutor to be useful in its content and format. However, as shown in Table 15, a desire for hybrid pronunciation classes emphasizes the potential of technology in regards to time and schedule flexibility, variety of and appeal of materials, and opportunity to revisit materials along with the need for human interaction.
Table 15. Speaker’s Favored Type of Pronunciation Instruction and Rationale

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Favored types of instruction</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Hybrid</td>
<td>“I think interesting materials can attract students' attention toward them.”</td>
</tr>
<tr>
<td>S2</td>
<td>Hybrid</td>
<td>“I will have face to face exercise but will use the software often.”</td>
</tr>
</tbody>
</table>
| S3      | Fully online                 | “1. Flexibility  
2. A lot of interesting activities (watching a comic movie and learn)  
3. Using online software  
4. Stop working on the software whenever I want and then continue when I am ready” |
| S4      | Fully online                 | “I usually don't learn that much in face-to-face class[es] for some courses.” |
| S5      | Hybrid                       | “Hybrid so I get the chance to do more work online, plus have the chance to practice and ask question to a professor in the mean time and when needed.” |
| S6      | Fully online                 | No response provided. |
| S7      | Fully online                 | “This tutor is very amazing and it includes everything that is required to teach pronunciation, rhythm, intonation. Therefore, I don't feel like it should change.” |
| S8      | Fully online/ Hybrid         | “For me when some one goes to the regular classes they will not have the ability to extremely focus on the material, but when he has a chance he can watch the videos and focus as much as he could. In addition, I chose the hybrid to make a presentation maybe for some volunteered undergrads to evaluate this TA before and after taking this course.  
Also when someone take[s] the online class he can listen to every single word the lecturer says, and he can watch the video[s] as many times as he wish[es] and this give[s] him the ability to understand the material and get the point very well.” |
| S9      | Hybrid                       | “Short, straight-forward videos are more attractive than text!” |
| S10     | Fully online                 | “[the] Online tutor works as well as face-to-face teacher[s]. I can study by myself, and I can schedule it myself. [the] Online tutor is helpful and convenient.” |
| S11     | Hybrid                       | “I like to watch lectures and do exercises online, but I hope I can interact with instructor and peers..” |
| S12     | Hybrid                       | “I think talk[ing] to a real person whose English is very good is the best way to improve English.” |

Speakers’ evaluation of the Supra Tutor

The previous section examined speakers’ perceptions of their own performance in the Supra Tutor. Following is a discussion of the speakers’ evaluation of the tutor. Most specifically,
this discussion focuses on three aspects: usefulness, level of interest, and quality of the materials. Those three factors, which were thought to be related to learner motivation, may have had an impact on how comprehensible the naïve raters found the speakers to be. A brief discussion of speakers’ suggestions for improvement of the tutor concludes this section.

When it comes to perceived usefulness of the Supra Tutor, the average was 1.58 (SD = .51). That is, all of the speakers found the tutor to be useful. As for level of interest, the trend was similar to that found for perceived usefulness. The speakers found the Supra Tutor to be either extremely interesting or interesting (M = 1.50 (SD = .52). In terms of clarity of instructions, the average was 1.33 (SD = .65). In other words, the speakers found the instructions included in the tutor to be very clear. Finally, the speakers assessed the Supra Tutor in terms of overall quality. The average was 1.58 (SD = .51), which means that they regarded the materials included in the tutor to be of good quality. Table 16 shows the descriptive statistics for speakers’ evaluation of the online tutor. Figure 17 shows speakers’ individual assessments of the level of usefulness, level of interest, and clarity of instructions in the Supra Tutor. Figure 18 provides speakers’ individual ratings of the overall quality of the Supra Tutor.

Table 16. Descriptive Statistics for Speakers’ Evaluation of the Supra Tutor

<table>
<thead>
<tr>
<th>Variable</th>
<th>Descriptors</th>
<th>Mean (N = 12)</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of usefulness</td>
<td>1 = extremely useful, 2 = useful, 3 = somewhat useful, 4 = not very useful, 5 = not at all useful</td>
<td>1.58</td>
<td>.51</td>
</tr>
<tr>
<td>Level of Interest</td>
<td>1 = extremely interesting, 2 = interesting, 3 = somewhat interesting, 4 = not very interesting, 5 = not at all interesting</td>
<td>1.50</td>
<td>.52</td>
</tr>
<tr>
<td>Clarity of instructions</td>
<td>1 = extremely clear, 5 = unclear</td>
<td>1.33</td>
<td>.65</td>
</tr>
<tr>
<td>Overall quality of the Supra Tutor</td>
<td>1 = excellent, 5 = poor</td>
<td>1.58</td>
<td>.51</td>
</tr>
</tbody>
</table>
In terms of usefulness of the Supra Tutor, the speakers commented on different aspects. For instance, Speaker 7 mentioned that “Due to this tutor, I am now aware of which syllable to
stress and which not to. I got to know that dictionaries use symbols to indicate primary and secondary stress for words.” Speaker 9 mentioned that, “Using TV series videos to understand about American English was very useful.” Speaker 3 highlighted that “The activities were applicable and interesting” and that he learned how to use two pieces of software (Audacity and Praat) and that he liked using Audacity to imitate native speakers. Speaker 2 also commented on the usefulness of using the software to improve his English and emphasized that, “I really liked the Praat software. I can see [myself] using it often after this week [intonation module] too. In fact, Audacity too.”

As for level of interest, speakers also pointed out a range of different aspects about the tutor that engaged them. Most of their level-of-interest comments were related to their favorite activities. Speaker 6 mentioned that, “I like most of the online tutor. For example, the sequence of everything, the child’s voice (I guess your family had a great time together while developing the class materials), the selected words and presentation. Also, some videos are so vivid!” Audacity and Praat were perceived to be very interesting. Eleven of the 12 speakers elected exercises using those two platforms as being one of their favorite activities. Instructional lectures, quizzes, and perception and production exercises were also perceived to be engaging. Table 17 displays speakers’ rationale for their favorite activities.
Table 17. Speakers’ Rationale for their Favorite Supra Tutor Activities

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Favorite Supra Tutor activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>“Recording my own sound was the most interesting part to me. For two important reasons. First, I could follow my job and see the progress. Second, I could analyze my sound and compare it with the correct one.”</td>
</tr>
<tr>
<td>S2</td>
<td>“Introduction to the software.”</td>
</tr>
<tr>
<td>S3</td>
<td>“I was interested in working with audacity software. Because you could listen to a conversation and then try to speak like native speakers. In this condition, if you do not know the rules you will indirectly learn like a baby. :-)”</td>
</tr>
<tr>
<td>S4</td>
<td>“The parts [that] include audio or video. I’m tired of any pdf lectures to learn some stuff. My performance is getting better through listening and watching.”</td>
</tr>
<tr>
<td>S5</td>
<td>“Recording myself and analyzing the waves. I finally understood about changes in pith to convey information in a sentence (for example how to properly end a question sentence).”</td>
</tr>
<tr>
<td>S6</td>
<td>“My favorite parts are doing the quiz and following your reading in audacity, really interesting. In the quiz, I found I am so bad at stress, rhythm and intonation; however, I was kind of excited. In the reading part, I heard my pronunciation is so weird, I never noticed that! Besides, I am feeling progress in the whole study procedure, which make[s] me happy.”</td>
</tr>
<tr>
<td>S7</td>
<td>“All activities were amazing. But, my favorite activity was imitating and recording the dialogues from the different video clips. Especially, Rhythm production exercise: King of Queens: Phone number rhythm. The main reason for this being my favorite is that I never got bored, even if I record my dialogues for 100 times. This activity was fun and made me realized that, if I follow the rhythm properly while speaking, it would sound more interesting to the listener.”</td>
</tr>
<tr>
<td>S8</td>
<td>“I liked the most the type of activity when I hear the sentence and fill in the blanks, I also liked recording the words I hear and then compare it with each other. The reason was that I liked to hear and see (the waves) of my speech and compare it with the original words and thus helped me to improve my performance.”</td>
</tr>
<tr>
<td>S9</td>
<td>“My favorite activity in the tutor was the production exercise because I was able to apply what I have learned from the lessons, videos, and lectures.”</td>
</tr>
<tr>
<td>S10</td>
<td>“I like them all. They are all truly interesting. And useful!”</td>
</tr>
<tr>
<td>S11</td>
<td>“Recording exercises. The recordings help me to visualize my pronunciation and give me a straightforward way to know my pronunciation.”</td>
</tr>
<tr>
<td>S12</td>
<td>“Recording your [model] pronunciation because I can find my mistakes and from the comparison with the native pronunciation, I know how to change my pronunciation.”</td>
</tr>
</tbody>
</table>

As discussed above, the speakers found the Supra Tutor to be useful, interesting, and of good quality. In other words, the tutor was successful in engaging the participants and holding
their interest until the completion of the training. In addition, the fact that four speakers showed significant comprehensibility improvement after only four weeks of training is a further testament to the tutor’s ability to motivate the speakers to follow the training to a successful conclusion.

The speakers also commented on their favorite materials and activities and their rationale for their choice. They were also prompted to express their opinions about aspects of the tutor that they liked the least and what could be changed to make the tutor more effective in the future. The number one least favorite activity was the reading materials because speakers felt that “listening to the lectures was more enjoyable” (Speaker 6). Thus, a suggestion for change was to replace the readings with instructional lectures (videos). Another suggestion was to include more production exercises. Some speakers felt that the number of activities included in the tutor was not adequate to allow them to practice the newly acquired knowledge. Another interesting suggestion, which reflects speakers’ choice of a hybrid pronunciation course in Table 15, was for “the teacher to meet their students once a week to help them improve. During the meeting, the teacher can test and correct the students’ pronunciation” (Speaker 6) or include a chat room in the tutor so that students can “discuss some issues or questions they have” (Speaker 3).

**Research Question 3: Contribution of Suprasegmentals to Comprehensibility Ratings**

Research Question 1 focused on naïve raters’ perceptions of ITAs’ comprehensibility before and after training. One limitation of RQ 1 is that the quantitative data do not provide explanations as to what features native listeners (naïve raters) focused on when judging the speech of the ITAs. As a matter of fact, although listener perception is key to the comprehensibility construct, research has yet to shed light on the dimensions that native listeners attend to when judging the comprehensibility of L2 speech (Isaacs & Trofimovich, 2012).
Previous research has shown that the reliability of comprehensibility ratings is uncertain, suggesting that the construct is not sufficiently delineated as to its components. According to Isaacs and Trofimovich (2012) research has not examined the linguistic factors that listeners attend to when assigning comprehensibility scores. The authors add that this is a shortcoming of existing scales, which confuse the constructs of intelligibility and comprehensibility because the descriptors are vague (e.g., TOEFL iBT and IELTS). The purpose of Research Question 3 was to shed light on this issue by investigating the cognitive processes that raters go through when assigning comprehensibility ratings to L2 speech.

Research Question 3 is based on the use of the language-specific scale by the trained raters and how they rated the speakers in terms of the language features that were included in the Supra Tutor (word stress, rhythm, and intonation) and on features that were not part of the training (consonants, vowels, fluency, grammar, and vocabulary). Although the language-specific scale used by the trained raters was a 9-point Likert scale (1 representing positive ratings, 9 representing negative ratings), for purposes of analysis the ratings were grouped in thirds (see Table 8 in Chapter 3), discriminating among levels of proficiency. That is, any feature rated from 1 to 3 was considered to be most proficient. Any feature that was rated from 4 to 6 was considered to be of intermediate proficiency. Those features rated from 7-9 were considered to be of low proficiency. This procedure is similar to that employed in previous research. For instance, Munro & Derwing (1995) define accent as being mild (1-3 on a 9-point Likert scale), moderate (4-6), or strong (7-9).

The purpose of this research question was two-fold: 1) to investigate whether the raters commented on the areas targeted by the tutor and whether these features were connected to improvement in speaker comprehensibility and, 2) to analyze which language features native
listeners attended to when judging L2 speech. The data for the analysis of this research question were derived from the ratings by six trained raters (quantitative) and from 72 verbal protocol files that were audio-recorded while the raters performed their rating of each speaker (qualitative).

As with the naïve raters, the trained raters did not rate the pre- and post-tests of the same speaker to avoid familiarity effects. Table 18 below shows the set up for data collection from the trained raters. As seen in the table, each group of raters rated 12 lectures in alternating order (pre and post-test).

**Table 18. Data Collection Design for the Trained Raters**

<table>
<thead>
<tr>
<th>Groups of trained raters</th>
<th>N</th>
<th>Lectures rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>3</td>
<td>Speaker 1 pre&lt;br&gt;Speaker 2 post&lt;br&gt;Speaker 3 pre&lt;br&gt;Speaker 4 post&lt;br&gt;Speaker 5 pre&lt;br&gt;Speaker 6 post&lt;br&gt;Speaker 7 pre&lt;br&gt;Speaker 8 post&lt;br&gt;Speaker 9 pre&lt;br&gt;Speaker 10 post&lt;br&gt;Speaker 11 pre&lt;br&gt;Speaker 12 post</td>
</tr>
<tr>
<td>Group 2</td>
<td>3</td>
<td>Speaker 1 post&lt;br&gt;Speaker 2 pre&lt;br&gt;Speaker 3 post&lt;br&gt;Speaker 4 pre&lt;br&gt;Speaker 5 post&lt;br&gt;Speaker 6 post&lt;br&gt;Speaker 7 post&lt;br&gt;Speaker 8 pre&lt;br&gt;Speaker 9 post&lt;br&gt;Speaker 10 pre&lt;br&gt;Speaker 11 post&lt;br&gt;Speaker 12 pre</td>
</tr>
</tbody>
</table>
Table 19 shows the average comprehensibility ratings each group of trained raters assigned to each speaker. Given that the number of raters is only three in each group, inferential statistics were not conducted. The standard deviation values in Table 18 show that the raters in each group rarely agreed on their comprehensibility scores as a group. That is, when the scale is analyzed in thirds (1-3, 4-6, 7-9), all three raters agreed in a few instances, generally in the case of more proficient speakers. More frequently, however, two raters agreed, but the third rater assigned a different rating. While this two to one ratio of agreement is consistent, it is noteworthy to observe that the pairing of raters in agreement varied depending upon the feature being rated for a given speaker (see Table 20). This finding is in line with previous research that indicates that although raters tend to assign similar overall comprehensibility ratings to a given speaker, they tend to focus on different language features when judging the comprehensibility of L2 speech (e.g., Isaacs & Trofimovich, 2012). This issue will be discussed further later in this section.

The standard deviations also indicated that raters tended to assign more similar comprehensibility ratings on speakers’ post-tests as compared to their pre-tests. Speaker 2, who received significantly better comprehensibility ratings on his post-test based on the naïve raters’ judgment, was assigned remarkably more positive post-test comprehensibility scores by the trained raters as well. The standard deviation values for Speaker 2’s pre and post-tests varied greatly (pre $SD = 2.08$, post $SD = .57$), which indicates that raters assigned more similar comprehensibility scores on his post-test. A similar trend is seen for Speaker 3 and Speaker 8, who also received significantly more positive post-test comprehensibility ratings from the naïve raters. For these two speakers, the raters agreed 100% on their post-test comprehensibility ratings ($SD = .00$).
Table 19. Descriptive Statistics of Pre and Post-test Comprehensibility Ratings by the Trained Raters

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Pre Test</th>
<th>Post Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>S1</td>
<td>3.66</td>
<td>1.15</td>
</tr>
<tr>
<td>S2</td>
<td>7.33</td>
<td>2.08</td>
</tr>
<tr>
<td>S3</td>
<td>3.33</td>
<td>1.15</td>
</tr>
<tr>
<td>S4</td>
<td>4.66</td>
<td>1.15</td>
</tr>
<tr>
<td>S5</td>
<td>2.00</td>
<td>.00</td>
</tr>
<tr>
<td>S6</td>
<td>6.33</td>
<td>1.52</td>
</tr>
<tr>
<td>S7</td>
<td>5.00</td>
<td>2.00</td>
</tr>
<tr>
<td>S8</td>
<td>3.66</td>
<td>2.08</td>
</tr>
<tr>
<td>S9</td>
<td>3.66</td>
<td>.57</td>
</tr>
<tr>
<td>S10</td>
<td>3.33</td>
<td>1.52</td>
</tr>
<tr>
<td>S11</td>
<td>3.00</td>
<td>1.00</td>
</tr>
<tr>
<td>S12</td>
<td>4.00</td>
<td>2.64</td>
</tr>
</tbody>
</table>

Another finding, based on the descriptive statistics presented in Table 19, is that naïve raters and trained raters tended to assign similar comprehensibility ratings to the speakers. In other words, they agreed on perceived comprehensibility before and after the training. The same speakers received more positive comprehensibility ratings on their post-tests by both naïve and trained raters. Also, the same speakers received slightly more negative comprehensibility ratings on their post-tests by both naïve and trained raters. The only exceptions were Speakers 4 and 11. While Speaker 4 received significantly more negative post-test comprehensibility ratings from the naïve raters (see Table 13), she received more positive scores from the trained raters on her post-test (pre $M = 4.66$ ($SD = 1.15$), post $M = 4.00$ ($SD = 1.73$). However, the larger standard deviation value for post-test indicates that the group of trained raters who rated Speaker 4’s post-test had a larger disagreement in terms of individual comprehensibility ratings. As for Speaker 11, she was assigned roughly the same pre and post-test comprehensibility ratings by the naïve raters (pre $M = 3.90$ ($SD = 1.81$), post $M = 3.95$ ($SD = 1.84$)). However, she received more positive comprehensibility ratings from the trained raters ($M = 3.00$ ($SD = 1.00$), post $M = 2.33$).
This finding may be explained by the fact that trained raters are better able to distinguish which linguistic features have an impact on their judgment of L2 speech than naïve raters (Isaacs & Thomson, 2009).

Although the descriptive statistics revealed trends, the average comprehensibility ratings did not tell the full tale about how comprehensible each trained rater found the speakers to be. Table 20 displays the raw comprehensibility ratings assigned to each speaker. Again, the data were analyzed based on three scale increments (1-3 high, 4-6 intermediate, and 7-9 low proficiency). Qualitative data collected through the verbal protocols employed during the rating process were examined to determine which language features each rater attended to when rating each speaker. As Isaacs and Trofimovich (2012) point out, “Raters weigh multiple factors when assessing L2 oral proficiency” (p. 480). That is, although there is often a consensus as to an overall comprehensibility level, raters arrive at this consensus from different perspectives. The two to one rater agreement ratio noted previously further supports this point.

Table 20. Raw Pre and Post Comprehensibility Ratings Assigned by the Trained Raters

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Pre-test comprehensibility ratings</th>
<th>Post-test comprehensibility ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rater 1</td>
<td>Rater 2</td>
</tr>
<tr>
<td>S1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>S2</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>S3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>S4</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>S5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>S6</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>S7</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>S8</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>S9</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>S10</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>S11</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>S12</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 21 provides an average of pre and post-test ratings for word stress, rhythm, voice range, and intonation. The ratings for rhythm are an average of ratings for “stressed words in a sentence” and “unstressed syllables/words in a sentence” since these two features are responsible for rhythm. That is, English rhythm is created by the alternation between stressed and unstressed syllables. Also, “voice range” and “final intonation” are reported separately. Although voice range (pitch range) is a feature of intonation, they seemed to be perceived as two separate features by the trained raters.

Table 21. Descriptive Statistics of Pre and Post Ratings of Suprasegmental Features

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Word stress (N = 3)</th>
<th>Rhythm (N = 3)</th>
<th>Voice range (N = 3)</th>
<th>Final intonation (N = 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre Mean (SD)</td>
<td>Post Mean (SD)</td>
<td>Pre Mean (SD)</td>
<td>Post Mean (SD)</td>
</tr>
<tr>
<td>S1</td>
<td>4.66 (2.08)</td>
<td>4.33 (2.30)</td>
<td>3.16 (.75)</td>
<td>5.66 (2.25)</td>
</tr>
<tr>
<td>S2</td>
<td>6.66 (1.52)</td>
<td>5.66 (1.52)</td>
<td>7.00 (1.78)</td>
<td>6.16 (1.72)</td>
</tr>
<tr>
<td>S3</td>
<td>4.66 (1.52)</td>
<td>4.33 (.57)</td>
<td>3.33 (1.03)</td>
<td>2.00 (1.09)</td>
</tr>
<tr>
<td>S4</td>
<td>6.33 (2.88)</td>
<td>6.00 (1.00)</td>
<td>5.50 (3.01)</td>
<td>4.83 (1.32)</td>
</tr>
<tr>
<td>S5</td>
<td>2.33 (.57)</td>
<td>2.66 (2.88)</td>
<td>2.00 (.00)</td>
<td>1.83 (1.16)</td>
</tr>
<tr>
<td>S6</td>
<td>6.00 (2.64)</td>
<td>7.00 (1.00)</td>
<td>6.16 (1.72)</td>
<td>7.33 (.51)</td>
</tr>
<tr>
<td>S7</td>
<td>6.00 (3.46)</td>
<td>4.00 (1.00)</td>
<td>5.00 (2.36)</td>
<td>3.16 (.40)</td>
</tr>
<tr>
<td>S8</td>
<td>4.33 (1.52)</td>
<td>5.33 (2.08)</td>
<td>3.50 (1.37)</td>
<td>4.00 (1.09)</td>
</tr>
<tr>
<td>S9</td>
<td>3.66 (.57)</td>
<td>4.66 (1.52)</td>
<td>3.66 (1.03)</td>
<td>2.83 (1.60)</td>
</tr>
<tr>
<td>S10</td>
<td>4.33 (2.51)</td>
<td>5.00 (1.00)</td>
<td>4.00 (2.52)</td>
<td>5.33 (.81)</td>
</tr>
<tr>
<td>S11</td>
<td>3.33 (.57)</td>
<td>4.66 (3.78)</td>
<td>4.33 (.81)</td>
<td>3.33 (2.25)</td>
</tr>
<tr>
<td>S12</td>
<td>5.66 (1.15)</td>
<td>5.33 (1.15)</td>
<td>3.66 (2.42)</td>
<td>4.66 (.51)</td>
</tr>
</tbody>
</table>

Note. Rhythm = combined ratings for stressed words in a sentence and unstressed syllables and words in a sentence

Based on the means and standard deviations displayed in Table 21, a few conclusions can be drawn. First, the large standard deviation values across the table indicate that the trained raters weighed the suprasegmentals in different ways. For instance, the standard deviation value of 3.78 for speaker 11’s word stress post-test ($M = 4.66$) shows a large disagreement between the three trained raters, which was not the case of the group of raters who rated her pre-test ($M = 3.33$ (SD = .57)). Second, there was not a clear connection between comprehensibility ratings and the
ratings for the suprasegmental features targeted by the Supra Tutor. For instance, the only case in which a clear connection is seen is Speaker 3, who showed significant comprehensibility improvement based on the naïve raters’ perceptions (RQ 1). Speaker 3 also received more positive post-test comprehensibility ratings from the trained raters as shown in Table 18. Speaker 3 was assigned more positive ratings on all four suprasegmental features (word stress, rhythm, voice range, and intonation). Another case that indicates the lack of connection between comprehensibility ratings and suprasegmental ratings is that of Speaker 8. Although Speaker 8 received significantly more positive comprehensibility ratings on his post-test based on both naïve and trained raters, he was assigned more negative ratings on word stress, rhythm, voice range, and intonation, especially in regards to voice range (pre $M = 3.00$ ($SD = 2.64$), post $M = 5.00$ ($SD = 5.00$). On the other hand, Speaker 7 did not show improvement based on naïve and trained raters’ perception of her comprehensibility. However, except for final intonation, she was assigned more positive suprasegmental ratings (word stress, rhythm, and voice range) on her post-test.

Following is a discussion of the analysis of the verbal protocols regarding trained raters ratings of suprasegmental features. As stated previously, for the purposes of analysis the language-specific scale was analyzed in three increments (1-3 high, 4-6 intermediate, and 7-9 low) distinguishing speakers’ proficiency in the employment of each of the suprasegmental features.

As seen in Table 22, the raters appeared to have given roughly equal consideration to each of the analyzed suprasegmental features. The trained raters used the lowest ratings (7-9) sparingly, suggesting that they rarely found the speech to be incomprehensible because of the features targeted in the Supra Tutor. However, word stress, with a total of 81 observations,
seemed to be a feature that raters focused on most when assigning low proficiency values (19 occurrences). That means that word stress was perceived to have a large impact on the comprehensibility of L2 speech. This finding aligns with previous research on the impact of word stress on comprehensibility (e.g., Isaacs & Trofimovich, 2012; Murphy, 2004; Tanner & Landon, 2009; Zielinski, 2008).

**Table 22.** Frequency of Coded Suprasegmental Categories from Verbal Protocols by Proficiency Level

<table>
<thead>
<tr>
<th>Codes</th>
<th>High (1-3)</th>
<th>Intermediate (4-6)</th>
<th>Low (7-9)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word stress</td>
<td>21</td>
<td>41</td>
<td>19</td>
<td>81</td>
</tr>
<tr>
<td>Stressed words in a sentence (rhythm)</td>
<td>30</td>
<td>32</td>
<td>11</td>
<td>73</td>
</tr>
<tr>
<td>Unstressed syllables/words in a sentence (rhythm)</td>
<td>21</td>
<td>31</td>
<td>13</td>
<td>65</td>
</tr>
<tr>
<td>Voice range</td>
<td>38</td>
<td>25</td>
<td>11</td>
<td>74</td>
</tr>
<tr>
<td>Final intonation</td>
<td>30</td>
<td>35</td>
<td>6</td>
<td>71</td>
</tr>
</tbody>
</table>

In regards to word stress, raters generally commented on the frequency of misplacement of primary stress, equally distributed stress (i.e., equal stress on all syllables), and no obvious stress. In their explanation of their ratings, raters tended to express how word stress issues impacted their understanding of the speakers, which affected their ability to understand the lecture content. Based on the verbal protocol comments, word stress issues seemed to bother the raters more than the other features. This may be because, like segmentals, misplaced word stress is an identifiable language error. Misplaced rhythm and unexpected intonation are not as likely to be identified as errors. For instance, Rater 5, commenting about Speaker 4, stated “Okay. I guess for her the biggest thing I have written down as I look down my list is stress, stress, stress, stress, stress, so stress was an issue for me.” Table 23 provides samples of rater comments regarding word stress.
Table 23. Rater Comments Regarding Word Stress Issues

<table>
<thead>
<tr>
<th>Rater (R)</th>
<th>Speaker (S)</th>
<th>Issue noted</th>
<th>Rater comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>S1</td>
<td>Primary stress misplacement</td>
<td>“He had a couple words that I noticed, like, “CONstraints,” instead of, “conSTRAINTS,” and “PROduce,” instead of, “proDUCE.” Those were only two words that I noticed, so it’s not so bad, and those were I think words that could easily be confused with where the stress is in the word. I’m also going to give him a four for that.”</td>
</tr>
<tr>
<td>R2</td>
<td>S7</td>
<td>Primary stress misplacement</td>
<td>“I’m giving her an eight on this one because there was a lot of misplaced stress. This was a word she used a lot: instead of, “multiple,” she used, “mulTIple.” For the word, “hierarchical,” or, “hierarchy,” she said, “hierarchical,” so there was emphasis placed more towards the end of words. She also said, “approachED,” once instead of, “approached.””</td>
</tr>
<tr>
<td>R3</td>
<td>S6</td>
<td>Equally distributed stress</td>
<td>“The main stress on multisyllabic words: I think words like, “exaggerated.” “Exaggerated,” would be equally dispersed, the stress over all the syllables. “Broadcasted,” is equally dispersed again, so I gave her an eight on that.”</td>
</tr>
<tr>
<td>R4</td>
<td>S4</td>
<td>No obvious primary stress/Primary stress misplacement</td>
<td>“I would say when you take into account that sometimes it was unclear what the stress was, it’s a seven or an eight. “Often misplaced,” is probably a little extreme to say. Even thought there were only a couple of words that I felt were – oh yes, “recently,” too she pronounced, “reCENTly.” I guess three words. One was her own field. She said, “COMputer science,” and so that was really a problem. She said, “iMAges” or something. She was stressing that second syllable and she said – although she stressed, “validity,” correctly, she said, “validity,” and so that clear “a” which I’m sure she’s taking from, “valid,” did slow down processing.”</td>
</tr>
<tr>
<td>R5</td>
<td>S5</td>
<td>Primary stress misplacement</td>
<td>“His stress was off in longer words quite a bit. Words like, “development,” “pharmaceutical,” “organisms,” almost always those were off.”</td>
</tr>
<tr>
<td>R6</td>
<td>S7</td>
<td>Primary stress misplacement</td>
<td>“Main stress on multisyllabic words: I had problems with. I think there were several words and key words in the lecture I couldn’t understand because of misplaced stress. I can’t tell you what they are because I didn’t understand some of them. But some of the words I did understand I could make out the misplaced stress. Words like, “advanTages”; “advantages/advanTages.” “If your task is” – I think she was saying, “if your task is DISturbed.” Or maybe, “DIstill.” I don’t know what she was trying to say.”</td>
</tr>
</tbody>
</table>
As discussed previously, there was no clear connection between suprasegmental ratings and comprehensibility ratings. As for rhythm and intonation, once again the ratings showed that raters weighted the suprasegmentals in different ways. When it came to rhythm, three out of the four speakers (2, 3, and 9) who received significantly more positive comprehensibility ratings on their post-tests received more positive post-test rhythm ratings. However, that was not the case of Speaker 8, who also showed significant comprehensibility improvement. As a matter of fact, only three of the 12 ITAs did not receive more positive post-test rhythm ratings. All of the other nine ITAs were perceived to have improved their rhythm according to the trained raters’ judgment. For instance, Speaker 7, one of the ITAs that did not show significant improvement, received better rhythm scores on her post-test (pre $M = 5.00$ ($SD = 2.36$), post $M = 3.16$ ($SD = .40$). The smaller standard deviation (.40) seen for her post-test ratings indicates that the three raters who rated her post-test agreed on her perceived rhythm proficiency. A possible explanation for the improved rhythm ratings for nine of the ITAs is that the rhythm module in the Supra Tutor targeted different aspects of rhythm in both instructional materials and practice exercises. Also, the rhythm module seems to have been one of ITAs’ favorites because of the materials including sitcom scenes and songs.

As for intonation, two categories were coded: voice (pitch) range and final intonation. Although voice range is a feature of intonation, the trained raters seemed to have perceived them as two separate categories. As for final intonation, low proficiency ratings (7-9) were coded only six times, which indicates that final intonation was not a feature that impacted speaker comprehensibility. Voice range, on the other hand, was perceived to be an important feature for raters even though it did not cause problems with comprehensibility per se. That is, raters understood the speakers but mentioned that if they were students in that given ITA’s class, they
would have a hard time focusing on their lectures. As a matter of fact, the word “boring” was used countless times.

Table 24 displays the frequency of coded suprasegmentals per rater. Raters focused on different aspects when rating the proficiency of the speakers. This difference in rater perception highlights the need for a better definition of each suprasegmental feature and well-defined scale descriptors that distinguish among different levels of comprehensibility. Although the trained raters used in this study have linguistics backgrounds, it is possible that they do not have a clear understanding of the constructs being judged (Isaacs & Trofimovich, 2012). It may also be that they were not able to think phonologically in their judgments, since their specialist training as language teachers was not usually in pronunciation.

**Table 24. Frequency of Coded Suprasegmental Categories Per Rater**

<table>
<thead>
<tr>
<th>Codes (differentiated by proficiency level)</th>
<th>Raters (R)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
</tr>
<tr>
<td>Word stress 1-3</td>
<td>4</td>
</tr>
<tr>
<td>Word stress 4-6</td>
<td>6</td>
</tr>
<tr>
<td>Word stress 7-9</td>
<td>2</td>
</tr>
<tr>
<td>Stressed words in a sentence (rhythm) 1-3</td>
<td>2</td>
</tr>
<tr>
<td>Stressed words in a sentence (rhythm) 4-6</td>
<td>8</td>
</tr>
<tr>
<td>Stressed words in a sentence (rhythm) 7-9</td>
<td>1</td>
</tr>
<tr>
<td>Unstressed syllables/words in a sentence 1-3</td>
<td>4</td>
</tr>
<tr>
<td>Unstressed syllables/words in a sentence 4-6</td>
<td>5</td>
</tr>
<tr>
<td>Unstressed syllables/words in a sentence 7-9</td>
<td>1</td>
</tr>
<tr>
<td>Voice range 1-3</td>
<td>5</td>
</tr>
<tr>
<td>Voice range 4-6</td>
<td>9</td>
</tr>
<tr>
<td>Voice range 7-9</td>
<td>0</td>
</tr>
<tr>
<td>Final intonation 1-3</td>
<td>6</td>
</tr>
<tr>
<td>Final intonation 4-6</td>
<td>7</td>
</tr>
<tr>
<td>Final intonation 7-9</td>
<td>0</td>
</tr>
</tbody>
</table>
Analysis of the language features that trained raters attended to when evaluating the speakers’ comprehensibility (those features not emphasized in the Supra Tutor)

The previous section discussed raters’ perceptions of the suprasegmentals targeted by the Supra Tutor. When it comes to other language features (Table 25) that were not included in the online training, four findings stand out. First, an analysis of the means and standard deviations of the non-suprasegmental features leads to similar conclusions to those suggested by the ratings of suprasegmental features. The strongest agreement between the raters occurred with those speakers demonstrating the highest levels of proficiency (e.g., Speaker 5).

Second, vocabulary range and word choice and expression did not seem to have a negative impact on raters’ perceptions of speaker comprehensibility. However, the speakers delivered lectures on basic topics in their respective fields of study, which required them to possess a high degree of familiarity with the technical terms and expressions used during their presentations. In addition, the Supra Tutor included academic vocabulary in all four modules, especially in perception and production exercises. It is uncertain whether the results would be the same if the speakers delivered presentations on topics that were more specialized.

Third, pronunciation of vowels seemed to be the most consistent source of disagreement among raters. The large standard deviation values indicated that the perception of comprehensible vowel use varied widely among the raters. An obvious example of this divergence in vowel perception is seen in the case of Speaker 5. He was rated as the most proficient speaker before the training. The raters who rated his pre-test assigned him a vowel rating average of 2.66, with a standard deviation of .57, which indicates a general agreement. However, the raters who rated his post-test rated his vowel use at 3.33 (mean), with a standard deviation value of 2.30. Rater 1 gave him a 2, rater 2 gave him a 6, and rater 3 assigned him a
score of 2. Although his overall proficiency level remained in the top third, this is an illustration of how different raters react to different language features.

Finally, all of the four speakers who showed significant comprehensibility improvement (Speakers 2, 3, 8, and 9) received more positive fluency ratings on their post-tests (pre $M = 8.00$ ($SD = 1.73$), post $M = 6.00$ ($SD = 2.00$); pre $M = 3.33$ ($SD = 2.08$), post $M = 2.66$ ($SD = 1.15$); pre $M = 3.00$ ($SD = 1.00$), post $M = 2.00$ ($SD = .00$); and pre $M = 4.66$ ($SD = 2.51$), post $M = 2.66$ ($SD = 1.52$), respectively). This finding may be a positive outcome of using the Supra Tutor. Although fluency was not part of the training, it was a constant underlying feature in the activities, especially in the production exercises that required speakers to imitate a model. Another possible explanation is that the training allowed the speakers to be more confident in their oral abilities, which may be reflected in their post-training lecture delivery.
Table 25. Descriptive Statistics of Pre and Post Ratings of Non-suprasegmental Features

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Vowels (N = 3)</th>
<th>Consonants (N = 3)</th>
<th>Fluency (N = 3)</th>
<th>Grammar (N = 3)</th>
<th>Vocabulary range (N = 3)</th>
<th>Word Choice &amp; expression (N = 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre Mean (SD)</td>
<td>Post Mean (SD)</td>
<td>Pre Mean (SD)</td>
<td>Post Mean (SD)</td>
<td>Pre Mean (SD)</td>
<td>Post Mean (SD)</td>
</tr>
<tr>
<td>S1</td>
<td>3.00 (1.73)</td>
<td>6.33 (2.08)</td>
<td>3.67 (1.53)</td>
<td>4.00 (1.73)</td>
<td>3.00 (1.73)</td>
<td>4.66 (3.78)</td>
</tr>
<tr>
<td>S2</td>
<td>6.33 (1.52)</td>
<td>4.67 (.57)</td>
<td>4.66 (.57)</td>
<td>3.33 (1.15)</td>
<td>8.00 (1.73)</td>
<td>6.00 (2.00)</td>
</tr>
<tr>
<td>S3</td>
<td>5.00 (1.00)</td>
<td>5.00 (2.00)</td>
<td>3.33 (1.52)</td>
<td>3.33 (1.15)</td>
<td>3.33 (2.08)</td>
<td>2.66 (1.15)</td>
</tr>
<tr>
<td>S4</td>
<td>6.33 (3.05)</td>
<td>6.66 (1.52)</td>
<td>4.00 (1.00)</td>
<td>5.33 (.57)</td>
<td>2.33 (1.15)</td>
<td>3.33 (1.52)</td>
</tr>
<tr>
<td>S5</td>
<td>2.66 (.57)</td>
<td>3.33 (2.30)</td>
<td>2.33 (.57)</td>
<td>1.66 (.57)</td>
<td>2.00 (.00)</td>
<td>1.00 (.00)</td>
</tr>
<tr>
<td>S6</td>
<td>6.66 (.57)</td>
<td>6.33 (2.08)</td>
<td>6.66 (.57)</td>
<td>6.66 (1.15)</td>
<td>7.33 (1.52)</td>
<td>6.66 (.57)</td>
</tr>
<tr>
<td>S7</td>
<td>5.33 (2.51)</td>
<td>2.66 (1.52)</td>
<td>5.33 (2.88)</td>
<td>4.33 (2.30)</td>
<td>2.33 (.57)</td>
<td>3.00 (.00)</td>
</tr>
<tr>
<td>S8</td>
<td>6.33 (2.51)</td>
<td>4.00 (2.64)</td>
<td>3.66 (2.08)</td>
<td>3.00 (1.73)</td>
<td>3.00 (1.00)</td>
<td>2.00 (.00)</td>
</tr>
<tr>
<td>S9</td>
<td>4.66 (1.15)</td>
<td>5.66 (2.08)</td>
<td>4.33 (1.15)</td>
<td>4.33 (1.52)</td>
<td>4.66 (2.51)</td>
<td>2.66 (1.52)</td>
</tr>
<tr>
<td>S10</td>
<td>7.08 (2.00)</td>
<td>5.33 (2.08)</td>
<td>2.66 (1.52)</td>
<td>4.33 (2.08)</td>
<td>6.66 (2.30)</td>
<td>6.66 (1.57)</td>
</tr>
<tr>
<td>S11</td>
<td>2.66 (.57)</td>
<td>5.33 (3.21)</td>
<td>4.33 (.57)</td>
<td>6.00 (1.00)</td>
<td>2.33 (.57)</td>
<td>2.00 (1.00)</td>
</tr>
<tr>
<td>S12</td>
<td>5.66 (1.15)</td>
<td>5.66 (.57)</td>
<td>6.00 (2.64)</td>
<td>4.66 (1.52)</td>
<td>6.33 (2.51)</td>
<td>6.00 (1.00)</td>
</tr>
</tbody>
</table>

Note. Fluency = unnatural hesitations and pauses
Table 26 displays the frequency of coded non-suprasegmental categories by proficiency level. One apparent trend was that the use of vowels appeared to have had a large impact on raters’ perceptions of speaker comprehensibility. As seen in Table 26, vowel usage discriminated among levels of proficiency more than any of the other features. That is, speakers with vowel pronunciation issues were assigned to the low proficiency end of the spectrum. The vowel issues that the raters commented on were mispronunciation of vowels, vowel omission, and vowel insertion.

**Table 26.** Frequency of Coded Non-Suprasegmental Categories from Verbal Protocols by Proficiency Level

<table>
<thead>
<tr>
<th>Codes</th>
<th>High (1-3)</th>
<th>Intermediate (4-6)</th>
<th>Low (7-9)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowels</td>
<td>19</td>
<td>33</td>
<td>36</td>
<td>88</td>
</tr>
<tr>
<td>Consonants</td>
<td>31</td>
<td>48</td>
<td>16</td>
<td>95</td>
</tr>
<tr>
<td>Unnatural hesitations and pauses (fluency)</td>
<td>40</td>
<td>16</td>
<td>18</td>
<td>74</td>
</tr>
<tr>
<td>Grammar</td>
<td>35</td>
<td>34</td>
<td>20</td>
<td>89</td>
</tr>
<tr>
<td>Vocabulary range</td>
<td>39</td>
<td>22</td>
<td>3</td>
<td>64</td>
</tr>
<tr>
<td>Word choice and expression</td>
<td>45</td>
<td>25</td>
<td>5</td>
<td>75</td>
</tr>
</tbody>
</table>

Table 27 displays the frequency of coded non-suprasegmental features per rater. Once again, the raters seemed to focus on different aspects when rating the proficiency of the speakers. The vowel rating discrepancy noted above was especially true in the case of Rater 4 (21 observations). Table 28 illustrates the variety of features that raters focused on when judging the comprehensibility of the speakers.
Table 27. Frequency of Coded Non-Suprasegmental Categories Per Rater

<table>
<thead>
<tr>
<th>Codes (differentiated by comprehensibility level)</th>
<th>Raters (R)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
</tr>
<tr>
<td>Vowels 1-3</td>
<td>5</td>
</tr>
<tr>
<td>Vowels 4-6</td>
<td>8</td>
</tr>
<tr>
<td>Vowels 7-9</td>
<td>2</td>
</tr>
<tr>
<td>Consonants 1-3</td>
<td>5</td>
</tr>
<tr>
<td>Consonants 4-6</td>
<td>7</td>
</tr>
<tr>
<td>Consonants 7-9</td>
<td>0</td>
</tr>
<tr>
<td>Unnatural hesitations and pauses (fluency) 1-3</td>
<td>7</td>
</tr>
<tr>
<td>Unnatural hesitations and pauses (fluency) 4-6</td>
<td>3</td>
</tr>
<tr>
<td>Unnatural hesitations and pauses (fluency) 7-9</td>
<td>2</td>
</tr>
<tr>
<td>Grammar 1-3</td>
<td>5</td>
</tr>
<tr>
<td>Grammar 4-6</td>
<td>8</td>
</tr>
<tr>
<td>Grammar 7-9</td>
<td>1</td>
</tr>
<tr>
<td>Vocabulary range 1-3</td>
<td>5</td>
</tr>
<tr>
<td>Vocabulary range 4-6</td>
<td>3</td>
</tr>
<tr>
<td>Vocabulary range 7-9</td>
<td>0</td>
</tr>
<tr>
<td>Word choice and expression 1-3</td>
<td>7</td>
</tr>
<tr>
<td>Word choice and expression 4-6</td>
<td>5</td>
</tr>
<tr>
<td>Word choice and expression 7-9</td>
<td>0</td>
</tr>
</tbody>
</table>
### Table 28. Rater Comments on Other Language Features

<table>
<thead>
<tr>
<th>Rater (R)</th>
<th>Speaker (S)</th>
<th>Language Feature</th>
<th>Rater comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>S6</td>
<td>Fluency</td>
<td>“Hesitations and pauses: yes, very much so but not at a point where I was like, “come on, like, get your sentence or get your thought out.” It wasn’t at an eight or nine level, so a seven.”</td>
</tr>
<tr>
<td>R2</td>
<td>S2</td>
<td>Grammar</td>
<td>“Many that seem sophisticated,” but this was a singular, an example of a singular subject, so he should have used an S at the end of the word. I’m not sure if that’s an S deletion issue or that’s a grammatical issue – subject-word agreement.”</td>
</tr>
<tr>
<td>R3</td>
<td>S1</td>
<td>Vowels</td>
<td>“Okay. As usual when I do this I end up focusing mainly on vowels and stress of words and sentences and multisyllabic words.”</td>
</tr>
<tr>
<td>R3</td>
<td>S7</td>
<td>Consonants/ vowels</td>
<td>“Okay. Consonants: I’ll just start off with the words, “birthday party.” It took me a long time to figure out that she was talking about a birthday party. I think that is reflected in her difficulty with consonants and vowels.”</td>
</tr>
<tr>
<td>R4</td>
<td>S8</td>
<td>Vowels</td>
<td>“For vowels: he had said, “STOmach,” and, “BOORger,” instead of, “BURger.” For “good,” he’d said, “god,” or, “gad.” “Nervous,” sounded like, “nairvous.” “Chew,” sounded like, “chow,” or, “cho,”… He had a variety of vowel mispronunciations. I think they could hinder someone from understanding, because these are the stressed vowels that are getting – or the only vowel in the word, depending – that are being mispronounced. I’m going to put that at a nine… Kills people’s understanding.”</td>
</tr>
<tr>
<td>R5</td>
<td>S2</td>
<td>Vocabulary range</td>
<td>“His vocabulary range was pretty small. He did say some … but I think I’m going to give him a five. He didn’t really lead the students through anything with transitions or things like that so he could have done better in both of those.”</td>
</tr>
<tr>
<td>R2</td>
<td>S6</td>
<td>Word choice and expression</td>
<td>“Her vocabulary and word choice: I’m giving both probably sevens too. There were not particularly appropriate to what she was talking about, but again she used circumlocution strategies to try to compensate for that.”</td>
</tr>
</tbody>
</table>
When analyzing the raw data provided by trained raters’ verbal protocol results, no clear trends were apparent. As shown in tables 22, 24, 26, and 27, the trained raters’ assessments and comments regarding specific suprasegmental and non-suprasegmental scoring appear almost randomly distributed. While they agreed on the overall perceived comprehensibility of the speakers before and after the training, each of the trained raters appears to have focused on his or her ‘feature of choice.’ This means that even raters with linguistic background are unreliable in how they rate specific pronunciation features in speech. The next chapter, Chapter 5, further discusses this issue.

**Chapter Summary**

This chapter presented the findings of this research. With respect to Research Question 1: Impact of the Online Pronunciation Tutor on ITA Comprehensibility, statistical analysis showed that four out of twelve ITAs did in fact demonstrate significantly improved comprehensibility after using the tutor. Regarding Research Question 2: ITAs’ evaluation of the Supra Tutor, the chapter discussed the ITAs’ usage of the tutor and presented their evaluations confirming positive impressions of all aspects of the tutor's makeup and utility. Finally, in answer to Research Question 3: Contributions of suprasegmentals to comprehensibility ratings, the chapter presented the trained raters' evaluations of the speakers' comprehensibility and discussed which features seemed to have the greatest impact on those ratings. The chapter concluded with a discussion focusing on the effect that the three suprasegmental features had on the trained raters' assessments of comprehensibility.
CHAPTER 5: CONCLUSION

This mixed methods study investigated naïve raters’ evaluation of the comprehensibility of international teaching assistants (ITAs), using microteaching recordings from both before and after four weeks of online suprasegmental pronunciation instruction. To help determine if the online instruction was motivating to the participants, the study analyzed ITAs’ evaluation of the Supra Tutor in regards to usefulness, level of interest, and quality of the materials. Finally, the study examined whether trained language-teacher raters would notice the contributions of the suprasegmentals (word stress, rhythm, and intonation) in their comprehensibility ratings. The trained raters both rated comprehensibility and followed a think-aloud protocol in discussing why they made the comprehensibility ratings. This chapter summarizes the research findings. Next, the limitations of the study are explained. Finally, pedagogical implications and suggestions for future research are discussed, followed by concluding remarks.

Summary of Research Findings

Research Question 1: Impact of the Online Pronunciation Tutor on ITA comprehensibility

This question investigated whether naive raters (domestic undergraduate students) perceived changes in the comprehensibility of 12 ITAs as a result of their focused self-study using the Supra Tutor, a four-week fully online pronunciation tutor focusing on word stress, rhythm, and intonation. The 12 ITAs prepared and delivered lectures on basic topics in their respective fields of study. Each ITA delivered a lecture prior to the training and delivered another lecture after completion of the training. The 178 naïve raters then viewed and rated the lectures. According to independent samples t-tests, four speakers showed significant improvement in their comprehensibility ratings, one showed a significant post-test reduction in
comprehensibility rating, and the remaining seven speakers’ comprehensibility ratings remained unchanged.

Activity reports provided by the Supra Tutor were analyzed to help explain the quantitative findings. The four ITAs with improved ratings accessed the instructional materials multiple times. They completed all perception and production activities in the training modules, and all but one completed all review module activities. Additionally, all of them made use of all required software, including Audacity and Praat, thus receiving the aural and visual feedback crucial to pronunciation improvement. Finally, all four appeared to have taken note of their relatively low diagnostic quiz scores and revisited instructional materials while closely following activity instructions. Their improved performance review quizzes can be seen as evidence that they increased their awareness of English suprasegmental features.

The ITA with reduced post-test comprehensibility ratings did not take full advantage of the Supra Tutor. She viewed most of the instructional materials once or twice and completed the quizzes. However, she did not submit many of the production tasks and did not make use of either Audacity or Praat, thereby depriving herself of the feedback component of the Supra Tutor. For those production activities that she did complete, she submitted audio files recorded on her cell phone. The fact that she received more negative comprehensibility ratings on her post-test could indicate that the tutor raised her awareness of the suprasegmentals but she needed more time to put this awareness to work through more guided practice.

The ITAs with unchanged comprehensibility ratings appear, in some cases, to have suffered due to poor study habits or other external factors (e.g., time to dedicate themselves to the training). Many of these speakers received relatively low diagnostic quiz scores, but only accessed instructional materials once rather than taking advantage of the Supra Tutor’s ability to
allow for repeated practice as needed. Some of them completed several activities in each module on the same day, usually just before a given module’s due date, rather than spreading out the activities for short periods throughout the week, which would have increased the benefits of those activities. One of the ITAs already had relatively high pre-test comprehensibility ratings, so any incremental improvement over a four-week period may have been easily overlooked. One of the ITAs had grown up in an Indian English speaking environment and may have had greater difficulty in changing the way her English sounded because of her comfort with spoken English. Other ITAs did show evidence of active interaction with the Supra Tutor, following all instructions and completing all activities and quizzes, but did not show comprehensibility improvement. It is likely that these ITAs would also have benefited from additional time and practice.

As discussed previously, the data collection from naïve raters was conducted during 13 regular English 150 classes with the presence of the class instructor and the researcher. The data collection was tied to the English 150 curriculum in that the activity involved oral presentations/lectures delivered by the ITAs. After data collection, each instructor, in their own time, discussed the Do’s and Don’ts of oral presentations with their students based on their ratings of the ITAs. In addition to carefully planned data collection set-up and procedures (e.g., step-by-step data collection slides, rating calibration) to minimize issues with the raters, the presence and the guidance of each class instructor was essential to ensure that the raters performed the task as prompted. Except for a few students who arrived late to class, missing part of the procedures, those who were texting on their phones while watching the lectures, and a few who did not complete all six speaker evaluations, the naïve raters dedicated themselves to the task at hand. Most of the students seemed to find the activity interesting.
Research Question 2: ITAs’ evaluation of the Supra Tutor

This question analyzed ITAs’ perceptions of their performance in the training and their evaluations of the Supra Tutor in regards to usefulness, level of interest, and quality of the materials. The results indicated that the ITAs generally believed that they had been dedicated in their training. However, their perceptions did not always match their actual use of the tutor. In fact, three of the four ITAs who rated themselves as extremely dedicated were diligent in their usage as shown by the activity reports; however, the fourth ITA used her phone to record production tasks that made use of either Audacity or Praat or both. The majority of the ITAs rating themselves as dedicated were those noted to have accessed the materials only once, completing a week’s worth of activities in three hours. It is likely that some of the ITAs believed that accessing the materials at least once and performing all the activities, even if in three-hour blocks, meant being dedicated in the training. This is an indication that future research should educate learners about the importance of spreading out pronunciation work in order to maximize training effects. For instance, this education could have been part of the brief face-to-face instructions on how to navigate the Supra Tutor and complete the activities. Another strategy that may have been effective would be the inclusion of a short educational video at the beginning of the tutor’s first module to raise ITAs’ awareness of how pronunciation skills change and what needs to be done in order for them to make their own learning more effective.

All of the ITAs perceived themselves as taking full advantage of the tutor to learn about pronunciation, new technologies and new learning strategies. However, as previously noted, some failed to take advantage of the opportunity to revisit needed instruction and activities or the feedback provided through the use of Audacity and Praat.
All indicated that they believed that they learned more, or much more, using the tutor than they believe they would have learned in a face-to-face class. They all also believed that an online component would be preferred for future pronunciation classes; seven indicated a desire to take hybrid classes (online plus face-to-face instruction) and the remaining five would prefer to take fully online pronunciation classes. This is seen as an indication of ITAs’ appreciation for the scheduling flexibility, opportunity to revisit materials, and variety and appeal of materials provided by the tutor, combined with the need for human interaction and feedback. This is an indication that an interaction component should be included in the Supra Tutor. The purpose of this research was to investigate if fully online pronunciation instruction (absence of a teacher) can be effective. The positive findings in this study indicate that it can. However, for educational and future research purposes, an interaction component can be added to the tutor in two different ways. First, a chat room can be included in the tutor so that learners can interact with and give feedback to each other at all times. Second, synchronous online “office hours” with the teacher can be scheduled so that learners can use that opportunity to request feedback and or/clarifications. This online teacher-learner meeting could be scheduled once a week or more depending on the availability of the instructor. These virtual meetings are easily possible with tools such as Adobe Connect or Google Hangouts.

Another possibility is the use of the Supra Tutor in a hybrid course. For instance, students could do most of the work online (watching instructional materials and practicing through perception and production exercises) and attend class once a week to receive feedback from the instructor or to be assessed on their progress.

The ITAs’ evaluation of the tutor was examined through additional data gathered from the questionnaire. ITAs were asked to evaluate the level of usefulness, level of interest, clarity of
instructions, and overall quality of the Supra Tutor. They were also asked about their most and least favorite activities, most and least favorite aspects of the Supra Tutor itself, and suggestions for improvement of the tutor. All evaluations of the tutor were positive. ITAs found it useful, the activities and instruction interesting, and the instructions for use of the tutor and ancillary software clear. This was supported by their evaluation of the overall quality of the tutor, with all ITAs rating the tutor as excellent or very good. Finally, these ratings were further reinforced by the positive responses to the open-ended questions. For instance, in response to the question that prompted ITAs to suggest changes to the tutor, one ITA stated, “I don't think this tutor needs any kind of changes as it is fabulous and very unique.” When asked the question “What did you like the most about the online tutor?” one ITA mentioned that “In this online tutor, I liked the way it is conducted. Firstly, quizzes and puzzles are given to test your knowledge. Secondly, videos and materials are provided to teach you, and finally exercises to make you strong in that area. I was very amazed by the word stress production exercises. Another ITA provided the following response:

“Actually I liked many things in this online tutor, and this is because:

1. I can do the practices anytime I want, which makes me focus more on the material.

2. I had fun watching the comedy sections because it makes you laugh and this breaks the routine that we usually have in the classroom, and thus makes you excited to ask for more.

3. This online program gave me the ability to clearly differentiate between the native English accent and the other types of accents.

4. Variety of activities and practices were definitely very useful because it makes the program not boring and repeat itself as usually happens in the classroom.
5. One last thing is that, it gave me the ability to learn some new software and how to use them like Audacity and Praat.”

Research Question 3: Contributions of suprasegmentals to comprehensibility ratings

Six trained raters rated the ITAs’ pre and post-tests on features that were included in the Supra Tutor (word stress, rhythm, and intonation) and on features that were not part of the training (consonants, vowels, fluency, grammar, and vocabulary). Verbal protocols were employed during the ratings to shed light on the cognitive processes that native listeners go through while judging the speech of non-native speakers. These ratings and recordings provided the data analyzed to determine whether the raters commented on the areas targeted by the tutor and whether these features were connected to improvement in speaker comprehensibility. The data were also used to determine which language features native listeners attended to when judging L2 speech.

Raters tended to agree (evidenced by lower standard deviation values) more often on comprehensibility when rating post-tests than on pre-tests. Also, naïve and trained raters tended to agree on perceived comprehensibility on pre and post-tests in general. Consistently large standard deviation values for the evaluation of suprasegmental features showed that the trained raters evaluated suprasegmentals inconsistently. Indeed, there was no obvious connection between comprehensibility ratings and ratings for suprasegmental features, except for one case.

When analyzing the verbal protocol data, it appeared that the raters gave generally equal consideration to each of the analyzed suprasegmental features. Lower incidence of low ratings for suprasegmentals implies that the raters did not often find speech comprehensibility to suffer due to features targeted by the Supra Tutor. The exception to this was word stress, which was noted in several instances to have a large impact on L2 speech. A likely explanation is that, like
segmentals, word stress errors are easier for native listeners to identify than other suprasegmental features (rhythm and intonation).

Perception of features not included in the Supra Tutor also confirmed that raters were more likely to agree when rating ITAs with the higher proficiency levels. The segmental feature generating the most disagreement among the raters was vowel use. The training on suprasegmentals may have shown up in another feature, fluency. All four of the ITAs who showed significant comprehensibility improvement also received more positive ratings on fluency in their post-tests. In addition to those four ITAs, Speaker 5 also received more positive fluency ratings on his post-test (pre $M = 2.00$ ($SD = .00$), post $M = 1.00$ ($SD = .00$), with raters agreeing 100% on both pre and post-test ratings. Speakers 11 and 12 also received more positive post-test fluency ratings, but the difference between pre and post-test ratings was not very large (one third of a scale point). While fluency per se was not targeted in the training, enhancing fluency was an underlying principle in many of the Supra Tutor’s activities.

Limitations

This research contributes to the area of technology applied to pronunciation instruction as it is the first study to employ fully online pronunciation training through the use of a tutor designed to target specific pronunciation features and to address the needs of the audience (ITAs). While the findings indicate that online pronunciation can be effective and engage learners, there are three main limitations to this study. The first limitation concerns the absence of a control group. Given that this study is exploratory in nature and employs a within-subject design, the main objective was to develop and evaluate an online pronunciation tutor that can be effective in aiding international teaching assistants to improve their comprehensibility. This research project focused on the development and evaluation of the Supra Tutor, not on a
comparison between instructional environments (e.g., online versus traditional face-to-face instruction). The participants in the study used and assessed the effectiveness of the tutor through a pre and post-test design. Based on the findings in this study, the Supra Tutor will be fine-tuned and comparative studies may be conducted in the future.

A second limitation to this study is related to length of training. Despite the fact that the Supra Tutor was developed based on ITAs’ needs and that it took individual characteristics (e.g., intrinsic motivation, self regulation) into account, some learners may need more extended deliberate practice to improve their comprehensibility. The training was set to four weeks for the purposes of this study. This time frame restricted the quantity of perception and production exercises included in the training. The busyness of the ITAs was another limiting factor. Because of their graduate studies and teaching duties, a longer training and a large number of activities in each module may have made them feel that they could not take part, but the shorter time-frame may not have been sufficient to reveal progress.

The last limitation to this study is the absence of delayed post-tests to analyze knowledge retention and transfer to novel contexts. The post-training lectures were delivered in the week following the end of the training. That is, the speakers that showed significant comprehensibility improvement on their post-tests may have been able to apply the theory to practice. As for the speaker who received more negative comprehensibility ratings on her post-test and for the speakers who did not show improvement, the short time frame between completion of training and post-test may not have allowed them to process the gained knowledge and be able to put it into practice to the extent required for perceived improvement. The administration of delayed post-tests focusing on novel contexts would reveal whether participants retained the knowledge
acquired through the use of the Supra Tutor and whether they were able to transfer that knowledge to different contexts.

**Pedagogical Implications**

This is the first study to employ fully online pronunciation training in the service of improved comprehensibility. The Supra Tutor was developed to take into consideration major principles thought to enhance effective pronunciation instruction: access to multiple speech models, opportunities for large amounts of practice, a learning environment that mitigates anxiety, self-determined pace, and the incorporation of technological tools to enhance the learning experience. In addition to its assessment components (e.g., scored diagnostic quizzes and perception exercises), the tutor included multi-faceted instructional approaches and activities to help learners understand what the English suprasegmental features are and why they are so important in spoken discourse. In other words, learners were given explicit instruction and rules so that they became aware of these features and better able to monitor their progress through guided practice. This awareness was intended to allow them to transcend the teaching environment and monitor their speech in the “real world,” where they are unlikely to receive sufficient feedback on their oral communication skills.

Based on the findings in this study, the Supra Tutor was successful in engaging their interest throughout the training. The tutor was also perceived to be useful and of good quality. A few features that learners mentioned as being particularly beneficial about the tutor were flexibility, variety of materials, incorporation of a range of speech models, the use of *Audacity* and *Praat* for perception and production tasks, and the use of sitcom scenes from popular TV sitcoms (e.g., The Big Bang Theory) in “imitation” production tasks. The learners also
highlighted the fact that the Supra Tutor allowed them to revisit materials at their need and convenience, which is not often possible in the regular classroom.

One pedagogical implication of the Supra Tutor is that it was intentionally designed to make use of available off the shelf software and materials that can be acquired at little or no cost to educational institutions, instructors, and learners. Audacity and Praat, the two pieces of software used to provide learners with aural and visual feedback during the training are freely available to anyone with access to a computer and Internet connection. This means that teachers can design and develop activities that go beyond a “listen and repeat” approach to enrich students’ learning experience without the need for spending money or a huge amount of time creating effective and engaging activities.

Another pedagogical implication is that The Supra Tutor was designed to be adaptable with modules in addition to the suprasegmental instructional materials and perception and production activities presently included in the tutor. As shown in the analysis of verbal protocol data, raters placed high importance on the correct use of segmental features (i.e., vowels and consonants). Modules focusing on these features can be added to the tutor to further enhance learners' comprehensibility. The segmental modules could include the same types of instructional materials (e.g., video lectures) and activities (perception and production tasks) and add a “lip movement” segment to provide learners with visual illustration of manner of articulation of English segmentals.

The Supra Tutor concept is also easily adaptable for different audiences and contexts. It can be, for instance, adapted to provide pronunciation instruction to health care professionals working in English-speaking countries. The Supra Tutor would be especially useful in EFL contexts, where pronunciation instruction is scarce and where learners often do not have the
opportunity to practice English outside of the classroom. In this context, the tutor could function as an extension of the classroom so that learners can be exposed to large and varied amounts of target language input and practice imitating the input to improve their oral proficiency and pronunciation skills.

In addition to learner training, the Supra Tutor has potential for teacher training. As discussed previously, pronunciation is a crucial component of perceived spoken comprehensibility, especially in the case of adult learners. Pronunciation is often the first (and sometimes the only) thing noticed by listeners. It is also clear that advanced learners often seem to need explicit pronunciation instruction in order to improve. However, effective targeted pronunciation instruction is scarce for a range of reasons. One of the main reasons for this is related to teachers’ lack of training. Language teachers often are not equipped to address pronunciation properly in the classroom because of difficulties in establishing goals, difficulty in defining pedagogical priorities, and difficulty deciding on effective approaches to teaching (Derwing & Munro, 2005). This inadequacy of training leads teachers to often ignore pronunciation in favor of grammar and vocabulary, which are features that they feel more comfortable teaching. When teachers venture to include pronunciation in their language classes, they tend to rely on materials that are unlikely to address individual learner needs. The Supra Tutor can be employed as a training tool for teachers to raise their awareness of the form and function of pronunciation features, to provide them with ideas about how to teach the topics and to create more meaningful and effective perception and production exercises, and to give them self-confidence to embrace the instruction of pronunciation.

In addition, classroom instruction is often sensitive to national testing constraints. An online tool can allow teachers to make pronunciation instruction available without a lot of class
time being used. Given the slow pace of change, standardized tests are unlikely to include a
significant focus on pronunciation any time soon, and online instruction such as the Supra Tutor
offers a different way to approach pronunciation that many teachers and students may find
attractive

**Directions for Future Research**

The tutor designed and developed for this study focused on suprasegmentals (word stress,
rhythm, and intonation) based on previous research claims that suprasegmentals should be the
main focus of pronunciation instruction (Kang, 2010), especially for advanced adult learners.
Considering that suprasegmentals are more challenging to master (Busà, 2007) and are thought
to be more likely to impact speaker comprehensibility than segmentals (Isaacs, 2008; Jenkins,
2000), a focus on suprasegmentals is warranted. However, based on the findings in this study,
segmentals (consonants and vowels) seem to be a feature that native listeners focus on when
judging the speech of L2 speakers. Unlike suprasegmentals, which are more complex because
they may be related to functions of discourse (e.g., intonation), segmental errors are easier to
identify, especially errors related to the functional load principle (Derwing & Munro, 2006) and
to the addition (e.g., ismall) and omission of sounds (e.g., fi\textit{ve}). Future research could investigate
the impact of each category on speaker comprehensibility so that more clear guidelines can be
provided to instructors to aid them in selecting and prioritizing goals with emphasis on learners’
needs.

Another important aspect of this study was the finding that even raters with linguistic
training are inconsistent in how they judge the comprehensibility of non-native speech. Although
the raters in this study had linguistic training, not all of them seemed to have a deep
understanding of comprehensibility and of the features being judged, especially in regards to
suprasegmentals. The raters received brief training on how to use the language-specific scale and on the items on the scale. However, it was a misjudgment to assume that they could fully perform the task as it was required of them. A follow-up study in which the raters are better trained could be conducted to investigate whether a more structured training focusing on the features to be judged would make a difference in terms of inter-rater reliability. Another study that would likely yield interesting results would be to compare the rating consistency of trained raters with general linguistic background to that of a group of raters with pronunciation expertise.

Comprehensibility is a difficult construct to define and measure. Further research is needed to operationalize comprehensibility in a way that allows raters to be on “the same page” in regards to what it is that they are required to judge. This highlights the importance of reliable oral proficiency assessment. Pronunciation assessment is generally conducted for three main purposes: 1) to diagnose areas of weaknesses; 2) to assess achievement (whether a certain feature or features have been mastered); and 3) to assess language proficiency table.

As with pronunciation instruction, negotiating the conflicts of the nativeness principle and the intelligibility principle comprises a major challenge for effective pronunciation assessment (Harding, 2012). Although researchers advocate intelligibility as the primary goal for pronunciation teaching and assessment, existing standardized speaking tests often conflate the two principles by using the expression “native-like” when describing constructs in test descriptors (Harding, 2012). For instance, the highest level of the Cambridge Main Suite exam states that “pronunciation is easily understood and prosodic features are used effectively; many features, including pausing and hesitation, are native-like” (Harding, 2012, p. 5).
Another issue in pronunciation assessment has to do with how it is measured. Generally pronunciation is measured within the broader construct of speaking and no specific guidelines are given as to how certain pronunciation features are assessed. Moreover, descriptors sometimes do not distinguish between different levels of a given feature (e.g., fluency). For instance, in 2006 the ACTFL guidelines for foreign language proficiency described ten levels of proficiency; however, “the guidelines are strikingly random in describing how pronunciation contributes to speaking proficiency” (Levis, 2006, p. 245).

The shortcomings of current pronunciation assessment call for a better definition of the constructs being measured and more concrete descriptors of the features being assessed so that raters are better able to not only identify areas of weakness but also agree with each other on their assessment of a given speaker.

Better definition of constructs and descriptors would not only aid in the fair assessment of L2 speakers but also serve as a guiding principle for pronunciation instructors. In other words, these guidelines would provide teachers with a better understanding of the importance of each pronunciation feature and a basis to make informed decisions regarding their classroom practices and assessment of students.

Conclusion

There is no denying it; pronunciation instruction is a vital component in how listeners assess spoken English. Even when an L2 speaker’s grammar and vocabulary are outstanding, communication cannot take place below a certain level of pronunciation proficiency (Celce-Murcia & Goodwin, 1991). However, pronunciation instruction is often relegated, at best, to supporting role in EFL and ESL instruction. This study has shown that a focus on suprasegmentals may have a positive effect on speaker comprehensibility, as demonstrated by
the number of speakers showing comprehensibility improvement based on naïve and trained
raters’ judgments of their speech. The instruction however was not consistently effective, given
that the comprehensibility ratings of two-thirds of the participants were not perceived to have
improved. A further complication is that many EFL and ESL teachers are not trained to
recognize and understand the importance of suprasegmentals to comprehensibility. Even raters
with linguistics backgrounds sometimes rely upon the "I cannot define it, but I know it when I
hear it" rationale.

This study, like that of Isaacs and Trofimovich (2012), has revealed that segmentals also
play a role in how native listeners’ react to nonnative speech. Therefore, it is critical that both
segmentals and suprasegmentals be incorporated into the pronunciation instruction curriculum be
it in the regular classroom or in an online teaching environment. In light of the findings in this
study, it seems valid to claim that an online pronunciation tutor such as the Supra Tutor has the
potential to incorporate both features and provide learners with a meaningful and engaging
learning experience that targets their individual needs and provides them with the opportunity to
take charge of their own learning.
REFERENCES


APPENDIX A. ITA BACKGROUND QUESTIONNAIRE

Please answer the following questions carefully.

This section asks you to provide some personal information that will be essential in the analysis of the data and will greatly influence the results of the study.

NOTE: During data collection for this study, you will receive a fake number ID for confidentiality purposes; this ID will be used in the data analyses. The personal questions below are meant to provide the researcher with demographic data about the sample population being analyzed. Your identity will remain confidential at all times, and it cannot be determined based on the data collected for this study.

1. What is your full name?
2. How old are you?
3. What is your gender?
   ( ) Female ( ) Male
4. What is your home country?
5. What is your native language?
6. What is your current student status?
   ( ) MA student ( ) Ph.D student
7. How does being a teaching assistant benefit you?
8. Do you speak any foreign languages other than English? Which ones?
9. How many years have you been studying English?
10. Where did you study English before coming to the U.S.?

11. How much do you use English daily?
    ( ) 0-15 minutes
    ( ) 15-30 minutes
    ( ) 30-45 minutes
    ( ) 45-60 minutes
    ( ) More than 60 minutes

12. Please describe your daily use of English (e.g., where you use it; for what purpose; speaking and listening; just listening; etc.).

13. Have you ever had any kind of pronunciation instruction?
    ( ) Yes ( ) No
14. If yes, please describe what kind of pronunciation instruction you have had (types of content and activities, taught separately or combined with other skills class like listening, etc).

15. Please describe a situation in which your pronunciation was the cause of trouble with communication.

Thank you very much for completing this questionnaire!
APPENDIX B. NAÏVE RATER BACKGROUND QUESTIONNAIRE

Please answer these brief questions. This section asks you to provide some personal information that will be essential in the analysis of the data.

NOTE: During data collection for this study, you will receive a fake number ID for confidentiality purposes; this ID will be used in the data analyses. The personal questions below are meant to provide the researcher with demographic data about the sample population being analyzed. Your identity will remain confidential at all times, and it cannot be determined based on the data collected for this study.

1. Please provide the ID number assigned to you.
2. How old are you?
3. What is your gender?
   ( ) Male ( ) Female
4. Select the correct option:
   ( ) I am American.
   ( ) I am an international student.
5. If you are an international student, what is your country of origin?
6. What is your native language?
7. What is your educational background (e.g., I am a senior in Mechanical Engineer; I have a BA in Spanish)?
8. Do you speak any foreign languages other than English and/or your native language? Which ones?
9. What is your major?
10. If not American, how long have you lived in the United States?
11. How many international teaching assistants (ITAs) have you had classes with? You may count this semester as well.
   ( ) None     ( ) 1     ( ) 1    ( ) 3    ( ) 4    ( ) 5    ( ) 6    ( ) More than 6
APPENDIX C. NAÏVE RATER SCALE AND INSTRUCTIONS

Directions: Now that you have carefully watched and listened to the lecture, please rate the speaker according to the aspects below. Note that the scale ranges from 1 to 9, 1 being positive rating (e.g., Extremely fluent) and 9 negative rating (e.g., Extremely disfluent).

1. The speaker is

<table>
<thead>
<tr>
<th>Extremely easy to understand</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Impossible to understand</th>
</tr>
</thead>
</table>

2. The speaker speaks

<table>
<thead>
<tr>
<th>Quickly</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Very slowly</th>
</tr>
</thead>
</table>

3. The speaker has

<table>
<thead>
<tr>
<th>A native speaker accent</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>A very strong accent</th>
</tr>
</thead>
</table>

4. The speaker is

<table>
<thead>
<tr>
<th>Extremely fluent</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Extremely disfluent</th>
</tr>
</thead>
</table>

5. The speaker is

<table>
<thead>
<tr>
<th>A very good presenter</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>A very poor presenter</th>
</tr>
</thead>
</table>
# APPENDIX D. TRAINED RATER SCALE AND INSTRUCTIONS

Rater: ______________________ Rated: ______________________ Date: ____________

**Directions:** You will watch 12 seven-minute videotaped lectures on different academic topics. Rate each presentation based on the features below. Note that the scale ranges from 1 to 9, **1** being positive rating and **9** negative rating.

## 1. Consonants are always
- Clearly enunciated  
  - **1**  **2**  **3**  **4**  **5**  **6**  **7**  **8**  **9**  Poorly enunciated

## 2. Vowels are always
- Clearly enunciated  
  - **1**  **2**  **3**  **4**  **5**  **6**  **7**  **8**  **9**  Poorly enunciated

## 3. Main stress on multisyllabic words is
- Always placed correctly  
  - **1**  **2**  **3**  **4**  **5**  **6**  **7**  **8**  **9**  Often misplaced

## 4. Stressed words in a sentence are
- Always clearly pronounced  
  - **1**  **2**  **3**  **4**  **5**  **6**  **7**  **8**  **9**  Rarely pronounced clearly

## 5. Unstressed words/syllables are
- Always de-emphasized  
  - **1**  **2**  **3**  **4**  **5**  **6**  **7**  **8**  **9**  Rarely de-emphasized

## 6. The voice range is
- Lively  
  - **1**  **2**  **3**  **4**  **5**  **6**  **7**  **8**  **9**  Monotone

## 7. Final intonation
- Always sounds natural  
  - **1**  **2**  **3**  **4**  **5**  **6**  **7**  **8**  **9**  Rarely sounds natural

## 8. Unnatural hesitations and pauses
- Rarely occur  
  - **1**  **2**  **3**  **4**  **5**  **6**  **7**  **8**  **9**  Occur very often

## 9. The use of grammar is
- Always correct  
  - **1**  **2**  **3**  **4**  **5**  **6**  **7**  **8**  **9**  Rarely correct
10. Vocabulary range is

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Very limited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sufficient</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

11. Word choice and expression are

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Often inappropriate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always appropriate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

12. Overall, the speaker is

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Difficult to understand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy to understand</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Additional comments:
APPENDIX E. VERBAL PROTOCOL INSTRUCTIONS

1. You will receive a brief training on how to use the scale and on the items included in the scale.

2. A calibration will be conducted with one or two videos so that you can use the scale and check if you have any questions or concerns.

3. Each speaker should be rated independently. That is, speakers should not be compared to each other in terms of their speaking abilities because we are not rating them; we are rating each speaker separately.

4. Listen to each seven-minute lecture and take notes that will help you rate each speech sample.

5. Immediately after listening to a lecture and taking notes, you will rate that lecture “out loud” and be audio-recorded for later analysis.

6. Note that teaching skills and topic are not being rated. The focus of your rated is the speech itself.
APPENDIX F. ITAS’ EVALUATION OF THE SUPRA TUTOR QUESTIONNAIRE

Please answer the following questions carefully. Your feedback will help improve this pronunciation tutor for the benefit of future students.

1. Before you start evaluating the materials, let's evaluate your performance. As a student during this online unit, you were:
   (   ) Extremely dedicated
   (   ) Dedicated
   (   ) Somewhat dedicated
   (   ) Not very dedicated
   (   ) Not at all dedicated

2. As a student, I took advantage of the online materials to learn as much as I could about pronunciation, new technologies, and new learning strategies.
   (   ) Definitely true
   (   ) True
   (   ) Somewhat true
   (   ) Not at all true

3. In your opinion, how useful were the online materials?
   (   ) Extremely useful
   (   ) Useful
   (   ) Somewhat useful
   (   ) Not very useful
   (   ) Not at all useful

4. How interesting was the online tutor?
   (   ) Extremely interesting
   (   ) Interesting
   (   ) Somewhat interesting
( ) Not very interesting
( ) Not at all interesting

5. Thinking about the online tutor as a whole (not specific activities), how difficult was it?
( ) Extremely difficult
( ) Difficult
( ) Somewhat difficult
( ) Not very difficult
( ) Not at all difficult (a piece of cake!)

6. What was your favorite activity in the tutor? Why?

7. What was your least favorite activity in the tutor? Why?

8. On a scale of 1 to 5, how would you rate our online tutor in terms of quality (1 being excellent and 5 being poor)?
Excellent ( ) 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 Poor

9. In general, the instructions in the online tutor were:
Extremely clear ( ) 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 Unclear

10. Through the online materials, I _____.
( ) learned much more than I would in a regular classroom
( ) learned a little more than I would in a regular classroom
( ) learned as much as (same) I would in a regular classroom
( ) learned less than I would in a regular classroom
( ) learned much less than I would in a regular classroom

11. What did you like the most about the online tutor? You can mention more that one aspect (e.g. flexibility, variety of activities and practice, technology, etc). Give a detailed answer, please.
12. What were your least favorite aspects of the online tutor? You can write about more than one aspect. Mention everything that you did not like (from types of activities to problems with technology). Give a detailed answer, please.

13. As an international student yourself (and a nonnative speaker of English), what do you think are the most challenging aspects of online classes for these students?

14. Please give suggestions for changes to the tutor so that it will be more effective in the future.

15. Based on your experience with this online tutor, in the future you would take a pronunciation course_______. Please check all the options that apply.
   (   ) face-to-face (regular classes)
   (   ) completely online
   (   ) hybrid (half online and half in a regular classroom)

16. Please explain your choice (or choices) in the previous question (question 15). Give as many details as you can.

17. Please type any additional comments about the online tutor and your experience with it below.

Thank you very much for filling out this questionnaire!
APPENDIX G. SAMPLE OF HANDOUT GIVEN TO ITAS BEFORE EACH STUDY STAGE

STAGE 1: PRE-TRAINING LECTURE (Lima’s dissertation study)

BEFORE THE LECTURE

- You will receive the topic for the lecture face to face two days before the test.
- You will be given instructions regarding the format/delivery of the task.
- You will be given help on how to prepare for your lecture and suggestions on what to focus on during the presentation.
- You will receive a brief training on public speaking and classroom management skills and help on how to select and address main points included in the lecture material.
- On the day of the presentation, you will deliver your seven-minute lecture in a real classroom.
- Your lecture will be videotaped for rating.

DURING THE LECTURE

- **NO visual aid allowed:** it will be you, notes you may need, and the blackboard (or whiteboard).
- You may take 2 minutes to write any notes on the board before the 7-minute lecture starts.
- Pretend there is an audience there with you. Act like you’re looking at the students and talking to them. That will help you deliver a more natural lecture. 😊

AFTER THE LECTURE

- Right after the lecture, you will receive quick instructions on how to navigate the online tutor.
  - **Module 1 (Word stress):** becomes available on June 9 and should be completed by June 15.
  - **Module 2 (Rhythm):** becomes available on June 16 and should be completed by June 22.
  - **Module 3 (Intonation):** becomes available on June 23 and should be completed by June 29.
  - **Module 4 (Review):** becomes available on June 30 and should be completed by July 6.
Post-training lecture (Stage 3 and last stage) will be scheduled for July 9, 10, 11 through Doodle (an email containing the link will be sent to you. Same procedures for Stage 1 will be followed).

**Note: Please let me know at any time if you have questions or concerns!**

Reminder: Participant Rights

Participating in this study is voluntary. You may choose to stop participating at any time, for any reason, without penalty. However, you are encouraged to participate, as the main objective of this study is to help international teaching assistants like you to improve their spoken English. In order for this study to be successful, it is crucial that participants participate in all three stages of the study.
APPENDIX H. VERBAL PROTOCOL SAMPLE

Transcription ➔ Rater Four, Lecture One.

Speaker 1 [Researcher]: Rater four, lecture one.

Rater 4: Okay. I felt like this was a little bit on the boring side. He’s pretty monotone and so – I guess I’m not going in order here, but that was something that really stood out to me. He’s not horribly monotone but combined with – let’s see, so I guess I’ll give him a three. But combined with – no, three’s bad. Three’s good. I’m not giving him a three; I’ll give him a seven.

Combined with his thought groups, as a student I think I would – his thought groups are extremely broken. I would get really frustrated very fast and then I would just quit trying. His thought groups I’m afraid are at the bottom. I should be probably going in order but I guess I’m going in order of what caught my attention. No, that’s not thought groups.

Unnatural hesitations and pauses. Yeah, it’s a bad thing. He’s a nine, I’m afraid. Number one thing he needs to work on. He’s also slow. I don’t know that that really comes into this anywhere, but it does contribute to the boringness. His thought groups are broken and he’s slow and he’s not very lively. I think it would be just a few minutes and I would quit paying attention if I were his student, which would obviously hinder my understanding of the content.

He does have several grammar issues, but they don’t – none of them are things that are actually going to hinder anyone from understanding him. For example, he drops articles very consistently. He says, “mathematical model,” instead of, “a mathematical model,” or, “the mathematical model.” “Cannot have negative amount”: so there is no article. “We want to find optimal situation,” no “the optimal situation,” but I really don’t think any of those issues are likely to influence anybody’s understanding. He also says, “this region contain,” so that third person verb S. He drops plural S on occasion. But none of these things are particularly serious, so while I would say I will give him a seven – maybe I should give him a
six. I’ll give him a six for grammar. It’s not grammar he needs to work on to be more easily understood.

Just incidentally regarding word choice and expression I guess: his collocations are not always correct. He said, “The example that we made in a previous class,” and, “if we move at this direction.” Again by themselves neither of these things would hinder understanding but they do demonstrate that he hasn’t yet mastered these particular collocations, which are important for his field. I guess I would say, “often inappropriate”? Not really. Let me see. “Sometimes inappropriate,” I guess would be a five. Let’s see. Okay.

Vocabulary range is sufficient. I think his vocabulary range was fine. I didn’t notice any issues with his range so I’m going to give him one for that.

Overall the speaker is easy to understand/difficult to understand: it’s a hard thing to say actually. Let’s see. I guess I’d put him at a six. Part of it is he does actually have points that I haven’t talked about yet that he’s difficult to understand really. The other thing though is what I mentioned before, where you would lose motivation to try to understand, because of his monotone and slowness and all of that. I think he’s trying to be slow to be clear so it’s a sad thing where it’s actually I think having the opposite effect to what he wants.

Final intonation: sounding natural/not sounding natural/rarely sounding natural. I didn’t notice any issues with this final intonation. Maybe it was because I was caught up with the other things. I really have no clue. If I have to assign a number I’d rather assign something give him the benefit of the doubt, but I really don’t know so don’t count this as gold standard rating.

“Stressed words in a sentence are always clearly pronounced or rarely pronounced clearly,” – I’m sorry, not that one. “Main stress on multisyllabic words is always placed correctly or often misplaced.” The only major issue I saw where it was actually it was a situation where the word stress was misplaced was in the word, “constraints,” and he repeatedly put the stress on the first syllable. So he said,
“CONstraints,” and of course that affects the vowel quality. That was a big issue. I wouldn’t necessarily say his word stress was often misplaced except I can see where other people might say because he does omit some syllables that that affects-

Speaker 1: You mean a seven or you mean a three?

Rater 4: Oh, I mean a three. Good catch! Yes. I’m so glad you are listening and not just sitting doing email or something. Okay.

The biggest issue honestly that I think he needs – well, maybe not the biggest issue. There’s a couple of big issues, but one of the biggest issues I think he needs to address, given his field or at least this particular topic: he doesn’t know how to pronounce the English alphabet letters the way English speakers do. He pronounces X as, “ix,” and with visual support, probably students would adjust, but it could get – because it happens repeatedly, over and over, it could be something which becomes an irritation which would ultimately hinder their understanding. Without visual support they might not get it at all. Z sounded like “zi” or something. It was a relaxed vowel instead of a tense vowel. That also I think without visual support would be incomprehensible.

He definitely needs to work on the vowel choice – not choice, I think it’s not conscious – but his pronunciation of the vowels in the pronunciation of English letters. Also when he pronounced, “company,” he pronounced it, "campany," or something that has more of an A quality. His vowels sometimes do have issues. Because of that and because of the keyness of where he has his vowel issues: “company,” is a content word, and then obviously the letters that he was using in his equations were key to what he was trying to say – that’s a pretty serious issue. I guess I’m going to put that as an eight because this is something he needs to be working on.

He does omit syllables too, and so – “too,” I guess, compared to the previous speaker. “Parallel”: he said it I think a couple of times and his pronunciation of it was, "paral" or "para." I can’t even do it quite right. But it was missing that final L in
addition to having a syllable entirely omitted. Except for the visual support, I think I would not have really had any idea what he was saying. Also he repeatedly used the word, "feasible region," and I don’t think I understand probably until he was about two thirds through those times because he pronounced it as, "feusal region."

Based on that, I’m going to say that – oh, how am I going to mark this? “Unstressed words or syllables are always deemphasized/rarely deemphasized.” I feel like I don’t have a category for this, but I’m going to say they’re deemphasized but this is not positive. They’re so deemphasized they’re not there. I’m going to give it a nine because nine is bad, even though the descriptor doesn’t actually match it and I will write a note: “It was actually omitted!” I didn’t feel like I had quite the right category to cover that.

Besides, “constraints,” where he pronounced it as, “CONstraints,” he pronounced, "sixty," as, "sixteen," and, "thirty," as, “thirteen.” The N was very clear. This is not a mistake that someone who hasn’t learned anything makes. I think he has learned that there is a distinction in English between, “sixty,” and, “sixteen,” and how to make it, but he’s misapplying that. But obviously that has huge issues in people’s understanding and without the visual support there’s no way I would have even known he even made the mistake because it sounded very clearly like he was saying, “sixteen,” and, “thirteen.” I’m going to say this is very serious, so I’m going to give him an eight because that just seems like he’s making mistakes in places that are really bad.

Regarding consonants – regarding grammar, he did drop the S in, “five corner point,” so that plural S did get dropped there. I’m sorry: I’m sort of out of order.

Minor things for his consonants that I noticed: He initially pronounced, "we," as, "vee," I thought I heard; so that V replacement for the W. I thought there was an instance in which he had pronounced B instead of P and he pronounced TH instead of S but I didn’t note specific words where those were an issue and I don’t remember if they were particularly damaging.
I’m going to give him – I don’t know. Maybe a six. Where I know he is making consonant errors, I can’t really speak to how damaging those errors are. I think, even though I have gone terribly out of order, I have covered everything.