

The Effectiveness of Pronunciation Training Software in ESL Oral Fluency Development

BY

SZYMON ZUBEREK

M.A., University of Illinois at Chicago, 2015

B.A., University of Illinois at Urbana-Champaign, 2009

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master of Arts in Linguistics
in the Graduate College of the
University of Illinois at Chicago, 2016

Chicago, Illinois

Defense Committee:

Susanne Rott, Chair and Advisor
Kara Morgan-Short, Linguistics
Maja Grgurovic, Linguistics

I would like to dedicate this thesis to Rose Diskin, without whose continued support, encouragement and advice it would have never been finished.

ACKNOWLEDGMENTS

Dudley Field Malone, an American celebrity, once famously admitted that “never in his life has he learned anything from any man who agreed with him.” While in his case disagreement carried him to the heights of fame, in my case its consequences were much more modest. This thesis is a child of an argument I had with Dr. Mareike Müller, who at the time was the head of the basic language program in German at the University of Illinois at Chicago (UIC). As my supervisor, she would object to my ideas of implementing speech recognition software to help improve our students’ German pronunciation. For better or worse, her reservations, though well-supported by research, did not resonate with me. Convinced that work with software would help students improve their pronunciation, even if only providing them with additional oral practice, I started thinking about ways in which I could advance my argument. Naturally, the questions I had at the time only bred more questions later, fueling my curiosity, and ultimately making me shift concentrations from German to Applied Linguistics. This project is the culmination of everything I have accomplished during my time with UIC’s Department of Linguistics. Much like the work described therein, this thesis is a fruit of collaboration. It’s the result of insight, encouragement, and sacrifice from people willing to work close with me. I would like to thank Dr. Susanne Rott the head of my thesis committee and the director of this undertaking. Without her patience, guidance, and wisdom this project would have never advanced out of its early stages. Having patiently read, revised, and edited its numerous drafts, she lent this work immeasurable support. The current and future recognition of this thesis is hers as much as it is mine. The same should be said about the rest of the committee guiding this project. I have huddled the offices and crowded seminars of Maja Grgurovic and Kara Morgan-Short, wrestling with half-baked ideas. Rather than being dismissive, they patiently guided them towards conclusion. I have shared those seminars with some remarkable friends and colleagues. Becky Bonarek, Abby McMillin, Devin Ferreira, William Lewis, and Koral Daniel not only listened to me milling on about yet another project, but also shared their ideas on how it could be improved. Special shout-out goes to Bernie Issa, who together with Dr. Morgan-Short critiqued the very

early draft of this paper.

Special thanks also go to the colleagues at the International Teaching Assistant (ITA) Program, whose support and collaboration made this project possible. I would like to thank Vandana Loomba Loebel, the program coordinator, who not only patiently answered all my questions, but also trusted me with an unrestricted access to the program's digital archives and resources. Having endured my irregular schedule, accommodated more late research sessions than necessary, while keeping me well caffeinated – I truly could not have asked for a greater person with whom to coordinate my efforts. Of course, my work with the ITA Program would not have been possible without the help of Jenn Taylor, Kim Hansen, and Katie Sauers, who kept me company during those long, winter evenings.

At home, my partner Rose, has also been critical to this thesis. Every evening I spent buried in work, every rant she endured, and every statistic she has helped me churn out, have secured her role in this project. In truth, this thesis would not exist without her. Thanks also go to my brother Marcin whose unconditional support has not diminished over the years, and who is still my biggest advocate.

Last of course, is Dr. Müller, without whose unwavering resolve, brilliant organization, and iron discipline this project would have never taken shape. For as Gandhi once said, “Honest disagreement is often a good sign of progress.”

TABLE OF CONTENTS

<u>SECTION</u>	<u>PAGE</u>
I. INTRODUCTION.....	1
II. AUTOMATIC SPEECH RECOGNITION IN PRONUNCIATION TRAINING – ITS PROMISE AND LIMITATIONS.....	3
III. WHAT IS ORAL FLUENCY?.....	13
IV. HOW IS ORAL FLUENCY MEASURED?.....	19
V. PARTICIPANTS.....	35
VI. MATERIALS.....	39
A. Classroom Materials.....	39
1. Course Curriculum.....	39
2. Course Textbook.....	40
B. Online Materials.....	47
1. NativeAccent v.3.....	47
VII. PROCEDURE.....	53
A. NativeAccent – Scoring and Analysis.....	53
B. ITAD and the Final Presentations - Scoring and Analysis.....	55
1. ITAD.....	55
2. Final Presentations.....	56
3. Scoring and Analysis.....	57
VIII. RESULTS.....	59
IX. DISCUSSION.....	65
X. LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH.....	72
XI. CONCLUSION.....	77
CITED LITERATURE.....	79
APPENDICES.....	85
A. Appendix A.....	86
B. Appendix B.....	88
C. Appendix C.....	94
D. Appendix D.....	95
VITA.....	97

LIST OF TABLES

<u>TABLE</u>	<u>PAGE</u>
I. ORAL FLUENCY MEASURES EMPLOYED BY THE REVIEWED STUDIES.....	34
II. CONTROL GROUP.....	38
III. EXPERIMENTAL GROUP.....	38
IV. DESCRIPTIVE STATISTICS FOR THE CONTROL GROUP.....	60
V. DESCRIPTIVE STATISTICS FOR THE EXPERIMENTAL GROUP.....	60
VI. CHANGE IN FLUENCY VALUES WITHIN GROUPS (INITIAL VS. FINAL ASSESSMENT).....	61
VII. CHANGE IN FLUENCY VALUES BETWEEN GROUPS (INITIAL AND FINAL ASSESSMENT).....	62
VIII. TIME SPENT WITH NATIVE ACCENT V.3.....	63
IX. THE EFFECT OF THE TIME SPENT WITH NATIVEACCENT V.3 ON ORAL FLUENCY.....	63
X. CHANGE IN SPEECH RATE FOR THE CONTROL GROUP.....	66
XI. CHANGE IN SILENCES FOR THE CONTROL GROUP.....	66
XII. CHANGE IN SPEECH RATE FOR THE EXPERIMENTAL GROUP.....	67
XIII. CHANGE IN SILENCES FOR THE EXPERIMENTAL GROUP.....	68
XIV. FLUENCY IMPROVEMENT ACCORDING TO NATIVEACCENT V. 3	71

LIST OF ABBREVIATIONS

ASR	Automatic Speech Recognition
AH1	First group in the experiment by Garcia-Amaya and Lorenzo (2009)
AH2	Second group in the experiment by Garcia-Amaya and Lorenzo (2009)
AH3	Third group in the experiment by Garcia-Amaya and Lorenzo (2009)
AH4	Fourth group in the experiment by Garcia-Amaya and Lorenzo (2009)
CALL	Computer-Assisted Language Learning
CAPT	Computer-Assisted Pronunciation Training
DI	Discourse Intonation
ECITA	<i>English Communication for International Teaching Assistants</i> (2013)
EFL	English as a Foreign Language
ESL	English as a Second Language
ESL 401	English for International Teaching Assistants
FL	Foreign Language
IM	Immersion Program
ITA	International Teaching Assistant
ITAP	International Teaching Assistant Program
ITAD	International Teaching Assistant Diagnostic
L2	Second Language
NA	NativeAccent v.3
NL	Native Language
NNS	Non-Native Speaker
NS	Native Speaker
PF	Pronunciation Feedback
UIC	University of Illinois at Chicago

SUMMARY

The study examined the effectiveness of automatic speech recognition – pronunciation feedback (ASR-PF) software in oral fluency training. The software employed for this purpose was *NativeAccent v.3* - a pronunciation-training platform for the speakers of English as a second language. The experiment evaluated two groups of native speakers of Mandarin enrolled in various graduate programs at a large research university in the United States. All participants were enrolled in a course designed for international teaching assistants and emphasizing oral language production and pronunciation skills. Whereas the experimental group was exposed to *NativeAccent v.3*, the control group experienced no such exposure. The participants' fluency was assessed using the measures of speech rate and silent pausing. Automated fluency evaluations were performed at the beginning and at the end of the study using *Praat*. In addition, participants' performance was automatically evaluated by *NativeAccent v.3*. Due to its considerable limitations, the study showed modest influence of the software on the participants' L2 oral fluency development. Still, despite these limited effects ASR-PF software appears to be a promising teaching tool with a lot of pedagogical potential.

I. INTRODUCTION

One of the most exciting applications of automatic speech recognition (ASR) software can be found in language education. From *Rosetta Stone* to *Babbel* to *Duolingo*, today's speech recognition technologies offer unprecedented possibilities in facilitating language teaching and acquisition. Among their numerous uses, pronunciation training in both foreign language (FL) and second language (L2) contexts seems to be particularly well positioned to benefit from their advantages. According to Neri et al. (2003) ASR software offers additional learning time and material. It also provides an opportunity to practice the language in a stress-free environment in the comfort of student's home. With their ever-evolving capabilities, these systems also come with increasingly complex interactivity, whereby the computer is not only able to understand learner's speech but also to adequately react to it. In addition to carrying the potential to engage students in realistic language learning, ASR systems can also provide feedback on the quality of the learner's speech.

Doubtlessly, ASR software is an innovative technology that has already managed to redefine the field of Computer-assisted language learning (CALL). With advantages for learning extending far beyond classroom, it is only likely to further transform language education. With their modest beginnings in dictation and accessibility software, today's ASR programs are not only strikingly more effective than their clumsy predecessors, but many of them are also equipped with additional functions that turn them into valuable tools in language classrooms worldwide. It is these technologies that stand at the core of almost all of today's CAPT platforms. Outfitted with advanced pronunciation feedback (PF) modules and supplied with banks of engaging learning material these programs are marketed as the answer to the plight of overcrowded language classrooms. Yet, while all of them promise their users L2 pronunciation improvements, only a handful of them can be said to deliver on this promise.

The goal of this study is to examine the effectiveness of ASR-PF software in L2 pronunciation training. The software used to this end is *NativeAccent v.3* (NA) designed as a pronunciation-training platform for the speakers of English as a second language (ESL). The study takes a look at two groups of

native speakers (NS) of Mandarin enrolled in various graduate programs at a large, research university in the United States. The groups followed the same curriculum, fell within the same age brackets, and spoke the same native language (NL). The paper will open with an overview and assessment of the general effectiveness of CAPT software in ESL training. The discussion will be framed by the following two issues: how well do CAPT programs assess target-language pronunciation as compared to assessment performed by humans, and how effective these programs are in improving target language pronunciation in general. In addition to demonstrating the software's efficacy, this brief survey will point out that while CAPT technologies do generally improve L2 pronunciation, little is known about their impact on L2 oral fluency. NA attempts to address this shortcoming by focusing on oral fluency as one of its training areas. This emphasis on oral fluency as a key component of pronunciation puts the question of the effectiveness of CAPT software in L2 oral fluency training in the spotlight. Accordingly, the main objective of this study and the research question it attempts to address is: How effective is CAPT software (*NativeAccent v. 3*) in improving oral L2 fluency?

With the research question established, the thesis will proceed to define oral fluency. To that end a representative handful of studies reflecting the last few decades of research into oral fluency will be consulted. The picture of fluency emerging from this synthesis will be substantiated by a set of quantitative measures employed by the cited studies. The measures appearing most consistently across the reviewed research will be isolated and adopted as the benchmarks for the current project. With the quantitative measures clearly defined, the paper will then introduce both subject groups and describe the materials used in the experiment. Attention will be paid to the course curriculum as well as the software used. These sections will be followed by a detailed account of the procedure and the data that it yielded. The data will be processed, analyzed, and the outcomes interpreted in light of the recent research. The paper's final sections will be dedicated to a summary of the study and a comprehensive discussion of its various limitations.

II. AUTOMATIC SPEECH RECOGNITION IN PRONUNCIATION TRAINING – ITS PROMISE AND LIMITATIONS

Thanks to the continuing advances in speech recognition technologies, computer software is increasingly utilized in L2 pronunciation training. But, is its increasing popularity in any way indicative of its effectiveness? The current section will attempt to address this question by reviewing how ASR software can contribute to pronunciation training in non-native speakers in foreign and second language contexts. To that extent, it will take a look at its two central functions: the assessment of users' target language pronunciation and the overall effectiveness in pronunciation training. The former aspect will be considered in relation to pronunciation assessment performed by humans, while the latter aspect – being a function of the reliability of user pronunciation's automatic assessment – will be evaluated on the basis of pronunciation improvements observed in the participants. With these two dimensions of ASR software pronunciation training functionally interwoven, the research reviewed in this section will demonstrate how the two processes can work together to facilitate gains in target language pronunciation. The studies reviewed here were selected to represent a wide range of applications of speech recognition technologies in language education. Though they are not presented in a chronological order, they span roughly a half-century inquiry into that particular area of CALL, bringing to the fore advances in technology, pedagogy, and research. They were also chosen to reflect an array of languages taught and to highlight the unique challenges that come with each language. The third selection criterion was the effect of age on language learning. Lastly, each study reviewed here examines a different ASR platform used during language instruction.

Most of today's ASR-based CAPT systems employ two interrelated processes: speech recognition and feedback. Since the reliability of the former determines the accuracy of the latter, it is important to first take a look at how well these systems are able to assess non-native pronunciation. The study examining the reliability of *FluSpeak* performed by Kim et al. (2006) highlights the challenges that

come with designing a dependable ASR system. It does so by looking at the correlation coefficient between the pronunciation scores assigned by *FluSpeak* and those awarded by the NS raters.

FluSpeak is an ASR-based system designed for Korean learners of English as a foreign language (EFL). The program consists of four modules: pronunciation practice, intonation practice, dialogue expression practice, and a pronunciation test. All instructions are accompanied by animations demonstrating the works of the human speech apparatus during the pronunciation of the target sounds. Students can record their output for analysis and see their pronunciation pattern mapped on a spectrogram set against NS speech.

The study examined 36 Korean EFL students enrolled in a general English conversation course. In terms of their proficiency all subjects placed between the beginner and the intermediate levels. Following a twenty-minute warm-up, during which they were asked to familiarize themselves with the software, students had to listen to 15 sentences recorded by a NS.¹ After listening to the recordings, students were asked to produce their own reading of each sentence. They were allowed to record repeatedly, saving only the utterances they felt most confident about. When finished with the recording of individual sentences they were asked to read these same sentences as a connected discourse. The recordings were automatically rated for the pronunciation of individual words in a sentence as well as for suprasegmental intonation². As a result each student was assigned two distinct scores, which were then averaged and compared to the pronunciation scores assigned by the three NS raters.³

The low correlation value between the scores given by the native speakers and those assigned by *FluSpeak* pointed out that the feedback generated by the ASR program might not be reliable at all.⁴

¹ The text of each of these sentences was displayed on the screen.

² Suprasegmental or prosodic features refer to speech characteristics (such as tone, stress, and fluency) that accompany consonants and vowels. Not limited to individual sounds, these characteristics frequently apply to syllables, words, and entire phrases (Fox, 2000).

³ The raters listened to the recordings in random order, rating them on a scale from 1 to 4 for both pronunciation and intonation. The scores were then averaged and contrasted.

⁴ The authors explain that this surprisingly low correlation could be attributed the nature of the subjects' pitch in the target language. As the pitch patterns observed in non-native language learners tend to be somewhat unpredictable,

However, it might have also called attention to pitch differences among the learners.⁵ Even though it appears that the accuracy of recognition displayed by *FluSpeak* is quite low, one should keep in mind that this study is not absolute and its results should not be used as a base for generalizations. Moreover, it also does not tell us how effective *FluSpeak* actually is when it comes to improving pronunciation. Designed to determine whether *FluSpeak* can reliably assess target language pronunciation, the study clearly demonstrates that the software fails to do that. Yet it also points out that despite its low reliability scores, the software could still prove dependable when applying to larger subject populations and testing for shorter phrases – the two criteria that informed the design of the next two studies, one by Neri et al. (2006) and the other one by Precoda et al. (2000).

In both studies subjects were divided into three groups: the experimental group using the software with automatic feedback, the experimental group working with a version of the software without the feedback, and a control group exposed to traditional, teacher-fronted instruction (with neither software nor automated feedback). The studies expanded on the ideas motivating Kim's (2006) research. Rather than focusing on whole sentences, they employed short words and phrases. They also differed from Kim's (2006) experiment in that they featured larger subject samples. While different in details, both were very similar in terms of their design and objectives, which inspired their joint assessment. Whereas the study by Neri et al. (2006) focused on the effects of pronunciation training of Dutch taught to a multilingual group of immigrants into the Netherlands (second language context), the one by Precoda et al. (2000) examined the effects of pronunciation coaching of Spanish on a group of English NS in the United States (foreign language context).

the ASR software might have detected them as pronunciation mistakes. The other factor that might negatively influence the reliability score is the phrase length. To support this claim the authors cite a number of studies demonstrating the inverse relationship between the length and the correlation index. Considering that the phrases examined in this study are long, complex, and taking advantage of various intonation patterns, this explanation seems very plausible.

⁵ Since native speakers modeling the sentences speak with a pitch that varies naturally the reliability of the learner intonation score generated by the ASR software is expected to be low. Moreover, when assessing total pronunciation *FluSpeak* considers intonation a much larger portion of the score than the pronunciation of isolated phrases.

Let us first take a look at the latter study. The experiment by Precoda et al. (2000) examined 45 students with one to two years of Spanish FL instruction. All students were native speakers of US English, aged 19-54, and although five of them were not actively studying Spanish, others claimed to have been studying the language for five years prior to the experiment. The two experimental groups practiced in three, 30-minute sessions per week for the period of three weeks. Each student was recorded twice, first time at the beginning of the study, and then after approximately three weeks. For each recording session students were asked to read 53 sentences featuring traditionally problematic sounds disguised under semantically simple vocabulary. Mispronounced sentences were not recorded. The outcomes of this study were assessed in terms of two characteristics: the accuracy of learner pronunciation against that of a NS (expressed as log posterior probability values) and the overall speech rate (defined as the number of phonemes uttered per unit of time).

FreshTalk, the program used in that study was designed to automatically assess L2 reading pronunciation. The program features a large collection of texts judiciously modified for Spanish learners at the beginner college level. The practice begins with students selecting a fragment of text they are interested in practicing. Clicking on the text activates a recording modeled by a NS, which could either be played at the “normal” speed, or at a “slower, careful” rate. Following the model recording, students are asked to record their own reading – a process they can repeat as many times as needed. The recorded samples are then evaluated both by the automatic speech recognizer and the NS graders.

With regard to the automatic evaluations, students’ individual scores were displayed right next to their averages for the whole session. A bar graph showing a total number of scored words was also generated with each read phrase displayed in either red or green depending on whether it was below or above the “correctness threshold”. As far as NS assessment was concerned, all recordings were evaluated on a scale from one to six, where one corresponded with least, and six with most accurate pronunciation. The scores were then averaged and compared with the pronunciation ratings generated by the ASR-based recognizer. The conducted comparison revealed a high inter-rater reliability. The pronunciation

evaluation conducted by *FreshTalk* was in line with the scores assigned by human raters, demonstrating that, in contrast to the previous study by Kim et al. (2006), ASR software is able to reliably evaluate its users' target language pronunciation.

In addition to proving that ASR software could be used to reliably assess learners' pronunciation, the study by Precoda et al. (2000) has also verified the software's overall effectiveness in target language pronunciation training. When compared with the control group, students who interacted with *FreshTalk* experienced a small but statistically significant increase in their accuracy scores. They also reported enthusiasm for using the software in pronunciation training, with the first experimental group admitting that the feedback actually motivated them to use the system. However, the study also showed that despite the measurable improvement in pronunciation scores, the pronunciation rate of the two control groups did not increase as expected.⁶ The authors suggest a few explanations for this outcome. It is speculated that the same feedback might be more effective over a period of time longer than that of the study. The relatively small sample group might have also exaggerated the effect of the differences among the participants. Finally, it is also possible that the provided feedback was poorly designed.⁷

In contrast to beginner-Spanish *FreshTalk*, the ASR-based CAPT system featured in the study by Neri et al. (2006) was designed to provide pronunciation feedback in Dutch or English to the learners of Dutch as a second language. Written for learners of arbitrary NL backgrounds, the program offered a number of interactive activities designed to train the pronunciation of 11 Dutch phonemes that have consistently proven challenging to the learners of Dutch as a second or foreign language. The input featured in this study was phonetically very rich and delivered to the learners in both written and audio forms. Each recorded utterance was first checked by the automatic recognizer for semantic correctness and only then analyzed for pronunciation. If the program determined that the targeted phonemes were mispronounced, they were highlighted in red. A red, disappointed "smiley" face was also displayed,

⁶ There was no difference between the control and the experimental groups in terms of the speech rate. The two experimental groups did not differ in terms of the pronunciation rate either.

⁷ Student comments showed that the feedback design could be improved.

followed by a message informing of the specific pronunciation errors. In order to keep students motivated, the program displayed no more than 3 errors at a time. Students were expected to repeat the mispronounced utterances until they arrived at satisfactory pronunciation.

The goal of Neri et al. (2006) was to take a look at the following: the learners' appreciation of specific feedback, expert ratings on global segmental quality, and expert annotations on segmental errors. The study engaged 30 beginner learners of Dutch, coming from a variety of language backgrounds. All participants were divided in groups similar to those in the study by Precoda et al. (2000). The two experimental groups using the ASR software were asked to participate in one additional session per week lasting between 30 and 60 minutes. The whole study lasted four weeks.

Student utterances were recorded and randomly assessed by six expert Dutch NS raters on a 10-point scale. The main focus was on the segmental quality with other aspects such as word stress, sentence accent, and speech rate largely ignored. Much like in the study by Precoda et al. (2000), this experiment has also confirmed a high reliability of automated pronunciation assessment. Yet, perhaps more importantly, it also corroborated the assertion that work with ASR-based pronunciation training software improves target language pronunciation. Despite these exciting findings, it is important to point out that although segmental pronunciation quality improved for all three groups, the greatest improvement was shown by the group actually exposed to ASR feedback.⁸ In fact, the group's progress was so significant that its members were able to close in on the control group, whose pre-test scores were highest out of all three groups.⁹ Lastly the findings also demonstrated that students largely enjoyed working with the CAPT system, considering it useful in pronunciation practice.

If pronunciation is considered to be one of the most difficult aspects of speech to master by adult language learners, then how is it handled by children? The next study attempts to address this question.

⁸ This was established despite the fact that the difference in improvement in all three groups was found to be statistically insignificant. The authors claim that this was a result of small group sizes and a large variation in segmental pronunciation quality in each group as well as between the groups.

⁹ The post-test revealed a 7.6% error decrease in the experimental group receiving the ASR feedback, compared to only 1.4% decrease in the no-feedback, control group.

According to Pica (1994), early pronunciation training is highly encouraged by the opportunity to capitalize on the so-called, *critical period* in children's cognitive development. Moreover, early attendance to pronunciation issues has been shown to help prevent fossilization of bad pronunciation habits later in life (Pica, 1994). Still, one needs to keep in mind that the ease with which children acquire L2 pronunciation tends to be offset by the difficulties posed by the evaluation of their pronunciation patterns. Assessing L2 pronunciation in children is a lot more difficult than it is in adults as it shows a higher variability in acoustic performance than the developed adult speech.¹⁰

In their study, Neri et al. (2008) tried to determine if ASR-based pronunciation training can be effective in helping children improve their word-level pronunciation. To this end, the study examined two groups of 11-year-olds. The control group counting 15 participants received traditional, teacher-fronted pronunciation training, while the experimental group comprised of 13 participants was asked to work with ASR-based CAPT software called *Parling*. All 28 subjects were native speakers of Italian, attending the same school and participating in the same curriculum. All children shared one teacher, and were attending the same classes. They had all had four years of formal English training before the start of the experiment.

Parling, the program used in the study, was developed specifically for the purpose of word-level pronunciation training. It consists of several themed modules, each featuring a story, a story-based adaptive word game, and a set of targeted words. The correct pronunciation of those words has been modeled by English NSs. Word recordings were subject to evaluation by an ASR recognizer.¹¹

Before starting their training, children were asked to read and record a set of 28 isolated words, varying in respect to their articulatory difficulty, length, and frequency of occurrence.¹² Children were

¹⁰ Irregular pitch and stress patterns can prove especially problematic when assessing children speech with ASR-based CAPT software (Precoda et al., 2000).

¹¹ The program assesses pronunciation either as "correct" or "incorrect". More elaborate forms of feedback were found to be uninformative by children and their instructors.

¹² The words were adapted from one of the stories used in the training and selected to address the most commonly used British-English phonemes.

allowed to repeat the words until they felt satisfied with their own performance. The results were then rated by 3 British-English NS's on a 10-point scale, where 1 was assigned for the least, and 10 for the most accurate pronunciation. Each utterance was given 2 scores – one for the word and one for the speaker.

Not unlike Precoda et al. (2000) and Neri et al. (2006), the current study revealed a high correlation coefficient between the automatic pronunciation assessment and human rater scores. This of course lends further support to the idea that automatic pronunciation assessment is just as dependable as human ratings. The study also showed that both groups improved their pronunciation, and that the improvements were comparable. Notably, the pronunciation of difficult/unknown words improved substantially over that of easy/more common words, suggesting that training with ASR software may result in improvements comparable to those achieved by means of traditional, teacher-fronted instruction. These promising results are further boosted by the fact that the members of the experimental group had their training cut in half¹³, and that the feedback they received was greatly simplified. Had they been subject to the same conditions as their fellow pupils in the control group their performance would likely have been even better.

By examining small but representative samples of language learners of different backgrounds and across different age groups, the studies outlined above have shown that ASR-based CAPT programs are not only able to reliably recognize and evaluate L2 oral input, but that they can also generate meaningful pronunciation feedback, that may positively affect students' target language pronunciation. The types of feedback reviewed so far are simple, ranging from color-coding, through numeric scores and their graphic representations, all the way to written descriptions and humorous illustrations. Missing from this review is perhaps the most instructive type of feedback – videos and animations. The next study attempts to investigate this type of feedback and its impact on users' target language pronunciation.

¹³ They were only working for 30 minutes as compared to the regular 60-minute instruction provided for the control group.

Video and animated feedback was employed in Bernsen's study of pronunciation feedback for L2 learners of Danish (Bernsen et al., 2006). For the purpose of this study an ASR-based CAPT system was developed. The program focused on the pronunciation of individual Danish words.¹⁴ A total of 450 of the most commonly used Danish word tokens were selected based on their "phonetic richness", relevance for daily use, and their phonotactic combinations. Each word was displayed in Danish and accompanied by an English translation.¹⁵ Moreover, the Danish words that were more problematic, were also supplied with their simplified phonetic translations.¹⁶ In addition to accessing the original Danish words, their English translations, and occasional simplified phonetic transcriptions, learners could also listen to the native pronunciation, and watch a video showing a native speaker pronounce the said words. It is important to note that users were not required to review all of the targeted vocabulary and could record their pronunciation at any time. Each recording was awarded feedback on a scale from 0 to 2, where 0 meant the least, and 2 the most native-like pronunciation. The scores were accompanied by pertinent "smiley" faces. The program did not provide any corrective segmental feedback.

For the purpose of the study the system was installed in 9 language schools and a handful or similar institutions across Denmark. Subjects were asked to engage all 450 words in consecutive series within suitable intervals.¹⁷ Whereas some learners only trained selected words, others performed spells of extensive training in under a week, and still others barely did any training at all. This inconsistency was the main reason why only 22 out of 88 students participating in this study generated results suitable for further analysis.¹⁸ Critically, none of the learners managed to complete the entire training routine. The obtained data revealed that despite the numerous inconsistencies, and under averaged conditions, all

¹⁴ Full description of the system is available online. The document is in Danish and can be accessed at: <http://www.nis.sdu.dk/projects/CAPT/DUTManual.pdf>.

¹⁵ The authors note that English translations help learners understand the meaning of the target words.

¹⁶ The authors maintain that simplified phonetic transcription has proven very useful in studying the pronunciation patterns of Danish, which similarly to English are phonetically irregular.

¹⁷ Full training curriculum was assumed to be 10 series of 450 words over the course of 10 weeks.

¹⁸ These 22 students performed on average 19.1% of the curriculum with the terminal values recorded at 2% for the lowest and 43% for the highest completion rate.

students improved their pronunciation. What makes this study stand out is the revelation that video feedback proved to have a significant positive effect on pronunciation. Especially low and average-scoring students, all of whom have viewed pronunciation videos prior to recording their utterances, seemed to have benefited from this kind of pronunciation feedback. Hence, it might be concluded that even though not a single subject completed its curriculum, the study upholds the effectiveness of ASR-based CAPT software in L2 pronunciation training.

The studies outlined above demonstrate that ASR-based CAPT software is not only able to reliably recognize pronunciation patterns of L2 learners but that it can also accurately evaluate them. They also verify that ASR-based pronunciation software is an effective tool in improving users' target language pronunciation. However, as pointed out by Neri et al. (2008), improved pronunciation of individual sounds and words does not necessarily translate into improved pronunciation at sentence or discourse levels. While the reviewed software does well working with isolated words and phrases, it falls short on addressing the suprasegmental features of non-native speaker (NNS) pronunciation. Overall, there appears to be a need for applications capable of handling such suprasegmental features – a need that *NativeAccent* promises to address. The goal of this study is to investigate how well *NativeAccent* delivers on that promise. To this end, the study will attempt to answer the following research question: Did L2 learners' oral fluency improve as a result of their work with *NativeAccent*, and if so, how significant this improvement was?

III. WHAT IS ORAL FLUENCY?

Even though oral proficiency has been a prominent component of L2 instruction ever since the birth of audiolingualism, it was not until the late 1970s when linguists began to assess oral fluency as its separate component (Brown, 2007). One of the first scholars who critically approached the question of fluency was C.J. Fillmore. In his pioneering article titled “Individual differences in language ability and language behavior” (1979) Fillmore outlines oral fluency in terms of its four main components. The first of these components is the ability to talk at length with a minimum of pauses. This ability to fill time with talk may be illustrated with an example of sports announcers, recognized for their ability to douse each broadcast minute with a flood of words. The second pillar of fluent speech is the ability to bundle information into coherent, semantically dense sentences without having to resort to an excessive amount of linguistic fillers. Accomplishing it requires mastery of both the semantic and syntactic resources of a language. It is also essential that speakers who are fluent in this manner take care not to crowd the discourse with semantically vacuous material. This economy of expression prefaces the third pillar of fluency described by Fillmore as the ability to speak appropriately in different kinds of social contexts and situations. Adequately meeting an interlocutor’s social and communicative demands calls for an ability to find the right thing to say and to feel at ease saying it across an array of conversational settings. Such aptitude is associated with the ability to use language creatively and imaginatively – fluency’s fourth pillar. Fillmore illustrates such imaginative use of language with examples of expressing old ideas in new ways via linguistic devices such as humor, puns, metaphors, etc. He goes on to say that speakers who are fluent in this manner tend to create an impression of doing a very rapid pre-editing of what they say. Furthermore, they are also able to quickly sort through a range of alternative ways of responding to a situation and to choose the one that is most resonant or clever. In his conclusion, Fillmore points out that although these four pillars are clearly distinct from each other, gifted speakers know how to simultaneously employ them all.

Fluency defined as the ability to adequately address interlocutor's social and communicative demands across various socio-cultural contexts also stands at the core of Sajavaara's (1987) research. Sajavaara characterizes fluency as "the communicative acceptability of the speech act, or its 'communicative fit'" (p. 62). Of course, what is acceptable in a given communicative context will be dependent on a variety of contextual factors – an idea further explored by Rehbein (1987). Drawing on Fillmore's fourth aspect of fluency, Rehbein's own investigation points out that in fluent speakers "activities of planning and uttering can be executed nearly simultaneously" (p. 104). Similarly to Sajavaara, Rehbein also argues that fluency is a context-dependent phenomenon hinging on "the speaker's evaluation of the hearer's expectations" (p. 104). This understanding of fluency as a highly automatized and complex linguistic process is also present in the work of Schmidt (1992), who contends that "fluency in speech production is an automatic procedural skill", and that "fluent speech is automatic, not requiring much attention or effort" (p. 358). The above accounts of fluency are elegantly summarized by Lennon (2000) who describes it as a "rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language under the temporal constraints of on-line language processing."

This rather broad understanding of fluency was later refined by Skehan (2003) whose study focused on distinguishing among different kinds of oral fluency. His findings suggest that oral fluency could be divided into at least three different categories: speed fluency, breakdown fluency, and repair fluency. Speed fluency is associated with speech rate. Breakdown fluency is determined mainly by hesitations, and repair fluency is in this context a function of self-corrections, repetitions, reformulations, etc. – the linguistic tools speakers employ to monitor and manage "online" utterances (p. 266)¹⁹. Skehan points out that these three fluency types vary depending on the speech elicitation tasks used as well as their difficulty level (Tavakoli and Skehan, 2005). He also observes that combining these three fluency

¹⁹ For an overview of strategies used by second language speakers to maintain the comprehensibility and fluency of their oral production see Dörnyei and Kormos (1998)

types can assist in gaining a better understanding of the phenomenon. For instance, looking at speed and breakdown fluencies as aspects that are inherently interrelated helps one formulate the definition of temporal fluency – a language phenomenon encompassing measures such as the length of runs between pauses, speech rate, total amount of silence, total time spent speaking, the number of pauses, or their length. Thus defined, temporal fluency could be then employed as an oral proficiency indicator, reflecting one’s adeptness in real time speech organizing (Skehan, 2003, p. 258).

Luoma’s (2004) assessment of fluency takes a closer look at what Skehan would consider “repair fluency”. Luoma notices that fluency is characterized by “absence of undue hesitations and excessive pauses” – an observation that aligns her findings with the previously discussed research. Her work also points out that while pauses can be relatively easily quantified, the matter of assessing which ones are “undue” and which are “excessive” is more of a matter of subjective judgment than quantitative measurement, and thereby reveals more about the listener than the speaker (p. 88). Kormos (2006) concurs with Luoma on fluidity comprising the predominant feature of fluency. Similarly to Skehan, she also postulates that oral fluency is a complex phenomenon that should not be examined with only one method. For that reason she proposes four different approaches to defining the measures of oral fluency²⁰ (p. 162). Whereas the first approach is concerned with the temporal aspects of speech production, the second one combines the study of such temporal aspects with the interactive features of discourse (e.g. turn-taking mechanisms). She also notes that topic initiations, back channeling, substantive comments, latching, and overlapping are all discursive mechanisms that contribute to fluency judgments, even if to a limited extent.²¹ In her third approach, Kormos focuses primarily on the phonological aspects of fluency. She starts out with the observation that fluent speech is most often equated with connected speech, where

²⁰ Oral fluency is understood here as L2 learner’s speech.

²¹ This is confirmed by Riggensbach (1991), who found that topic-initiations, backchannels, substantive comments, latching and overlapping, as well as the amount of speech produced comprised but a limited contribution to fluency judgment.

certain phonological processes (e.g. consonant attraction) are at work²². It is the ability to produce connected discourse (i.e. the ability to speak in whole phrases instead of word-by-word utterances) that leads to the perception of fluent speech. Interestingly enough, she observes that the length of utterances, or the brevity of pauses, is far less consequential for the overall perception of fluency. Lastly, Kormos' fourth approach revolves around the analysis of formulaic speech and its role in the perception of oral fluency in L2 speakers. The approach is based on the finding that the development of L2 oral fluency hinges on two interrelated cognitive processes: the first one governing the use of prefabricated language units, also known as formulaic language, and the second one pertaining to automatization of encoding (p. 156). Kormos also notes that L2 learning in general, and automatization in particular, take place in a way that echoes Ullman's Declarative/Procedural model of language learning (Ullman, 2015). In conclusion, Kormos reassesses the temporal aspects of fluency.²³ In so doing she validates the findings of previous research claiming that when compared to the NS speech, the best quantitative predictors of fluency are the speech rate (number of syllables per minute), the mean length of runs (the average number of syllables produced in utterances between pauses), and the phonation-time ratio (the percentage of time spent speaking as a percentage proportion of the time taken to produce the speech sample).²⁴

The above accounts of fluency in L2 speech are reviewed and systematized by Segalowitz (2010). Segalowitz frames L2 fluency as a purely performative phenomenon involving fluid, movement-like acts of speech. He argues that quantifiable, L2 speech is impossible to characterize in absolute terms (p. 39). Consequently, all that may be said about fluency is that under a particular set of circumstances L2 speech

²² Kormos emphasizes that consonant attraction has been shown to be a reliable indicator of the fluency of nonnative speech in informal English.

²³ Her study employs the following temporal measures of fluency: speech rate, articulation rate, phonation-time ratio, mean length of runs, silent and filled pauses per minute, disfluencies per minute, and pace (Kormos, 2006, p. 163)

²⁴ In regards to pausing, frequencies of silent and filled pauses tend to distinguish between fluent and non-fluent speakers in studies with a small sample of participants (see Lennon, 1990 and Riggenbach, 1991). In studies with a higher number of participants the number of filled and unfilled pauses tend not to correlate with fluency ratings (Kormos, 2006). Lastly, disfluencies occur in clusters in NNS, while fluent speakers tend to pause at grammatical junctures (Lennon, 1990, Towell et al., 1996).

may retain certain objectively measurable characteristics, which the listener might interpret as either fluent or non-fluent in particular ways (ibid.). What follows is that there can be no direct relationship between L2 speech and the speech quality that might be referred to as fluency. This claim is further corroborated by the lack of compelling, consistent patterns of oral production that may be universally recognized as reliable fluency indicators (p. 41)²⁵.

If fluency can be neither quantified nor even isolated as a strategic component of L2 oral performance, is it then at all possible to define it? Segalowitz contends that although a workable definition of L2 fluency is possible it is essential to break the phenomenon down into three separate fluency types: cognitive fluency, utterance fluency, and perceived fluency (pp. 46-52). Cognitive fluency pertains to the efficiency of operation of the cognitive processes underlying the production of utterances. Utterance fluency relates to the auditory features of utterances that reflect speaker's cognitive fluency. Finally, perceived fluency relates to the inferences made by listeners about speakers' cognitive fluency based on their perceptions of the said speakers' utterance fluency. This tripartite understanding of oral fluency not only gracefully ties in prior research, but it is also promising for the future inquiry into the phenomenon. By concentrating on its three aspects – the efficiency of the underlying cognitive processes, the temporal and repair features of utterances, and their communicative acceptability – one is likely to arrive at more comprehensive understanding of fluency as compared to its reduced sense of a performative characteristic oscillating on a latitudinal continuum.

Segalowitz's account of oral fluency is a good point of departure for positioning fluency in the framework of the current project. Since this study does not approach any of the cognitive mechanisms informing oral fluency, the cognitive aspect of Segalowitz's definition will be left unaddressed.

Furthermore, as the study neither consults native speakers in evaluating the subjects' speech, nor does it

²⁵ According to Segalowitz fluency can be measured based on: individual observations/ratings, comparing L2 speech samples to similar L1 samples, and taking samples from the same speakers at different times during their L2 development. Adequate speech samples can be elicited via reading tasks, picture description tasks, story-retelling tasks, and spontaneous speech samples.

investigate NS perceptions of the subjects' fluency, Segalowitz's perceived fluency also will not have much bearing on our definition. The remaining dimension of oral fluency, described in terms of distinct, measurable auditory features of utterances, perfectly encapsulates the sense of oral fluency explored in this project. Hence, this paper will define oral fluency in terms of a set of quantifiable auditory features. Considering the multitude of such features, some of them will be more accurate in assessing oral fluency than others. In effort to determine which ones of these markers are most reliable, and suitable for this experiment, the next section will survey a sample of studies examining the degree of correlation between such features and their perception by human raters. The most frequently used features, that also consistently show high correlation with human rater scores will anchor the current study's definition of L2 oral fluency.

IV. HOW IS ORAL FLUENCY MEASURED?

Attempts at assessing speech fluency with temporal variables date back to the middle of the twentieth century (Goldman-Eisler, 1958). Already by mid-sixties oral fluency was well understood and the techniques employed to measure it had been developed (Goldman-Eisler, 1968). The majority of early research into the phenomenon was concerned mainly with examining the role of pauses in a monolingual context. These early studies were later elaborated on by researchers such as Grosjean and Deschamps (1972, 1975), who included comparisons between different languages. Pioneering cross-linguistic work was continued throughout the eighties by linguists such as Fathman (1980) and O'Connell (1980). Towards the end of the decade Raupach (1987) and Sajavaara (1987) extended their research on temporal variables by examining the language produced by second language learners. Raupach (1987) also attempted to situate his findings in the context of cognitive linguistics. His efforts should come as no surprise for while studying temporal variables for their own sake is important, integrating the obtained results into a model of production, perception, and language acquisition is even more valuable. Due to both their nature and consequence for the current project the other studies in this section have been presented in much greater detail. In effort to make their findings more transparent, and to facilitate their critical evaluation, they have all been summarized in TABLE I at the end of this section.

In his review of studies of speech production, Towell et al. (1996) isolated four main temporal variables used to analyze oral fluency: speaking rate, phonation/time ratio, articulation rate, and mean length of runs. The study assessed the development of fluency in advanced learners of French. The four-year long longitudinal experiment involved 12 students - four males and eight females – who were selected based on the uniformity of their university entrance examination scores. Upon the completion of the second year in the French language program, students were asked to watch a short French film. They were then asked to narrate it in French. Their narrations were recorded. The procedure was repeated at the end of the third year in the program and after the participants had returned from a yearlong study abroad trip, six months of which were spent in a French-speaking country. There were two recordings

made exactly one year apart. Students' fluency was assessed using the already mentioned four temporal variables. In terms of the mean length of runs it is important to mention that the pause cut-off point was set at 0.28 seconds.²⁶ The improvements were mainly accounted for by the increase in the mean length of run scores, which the authors single out as the best indicator of fluency development. Speaking rate, although arguably not as reliable an indicator as the mean length of runs, was also shown to measure fluency improvements.

The 1998 study by Strik, Boves, and Cucchiari, extends the above research by implementing an automatic speech recognizer in fluency examination. The experiment investigated reading fluency in 80 subjects: 20 native and 60 non-native speakers of English. The participants in the first group were heterogeneous in respect to region of origin and sex while those in the second group varied in terms of language background, English language proficiency level, and sex. All subjects were divided into three groups, and each group was assigned three raters. These nine raters (three phoneticians and two groups of three speech therapists) scored the recordings for fluency. The raters could listen to the recorded samples as many times as needed, and then rate them for fluency on a scale from 1 to 10. Each subject produced two sets of phonetically rich samples over the telephone. It took on average 30 seconds to read each sample. After being assessed by human raters, the recordings were analyzed by an automatic speech recognizer. Automatic speech assessment was performed based on the following measures: rate of speech, phonation-time ratio, articulation rate, total duration of sentence-internal pauses, average length of pauses, number of silent pauses, mean length of runs, number of filled pauses, and number of disfluencies (repetitions, restarts, and repairs). The results were analyzed to ascertain if significant differences could be observed between the two groups of subjects. Unsurprisingly, native speakers were shown to be considerably more fluent than their non-native counterparts. This difference in fluency is a

²⁶ Towell et al. (1996) point out that pausing is variable by task. While some pauses indicate that a speaker might be at the point where she or he is deciding what to say, others might emerge as a result of the speaker deciding how to articulate it (i.e. verbalize something she or he already has in mind). Thus it appears that pauses may reflect: demands of a particular task, learners' individual characteristics, the difficulties that individuals face when arriving at the right utterance, as well as those experienced when verbalizing those utterances.

function of two factors: fewer pauses and disfluencies as compared to non-native speakers, and the speed of delivery higher than in non-native speech. With the exception of the average length of pauses, all tempo-related variables were strongly correlated with fluency ratings. Hesitation phenomena such as filled pauses and disfluencies demonstrated no strong correlation with fluency scores.²⁷ The study revealed that quantitative variables such as rate of speech, phonation/time ratio, number of pauses, and mean length of runs are able to predict fluency scores with a high degree of accuracy. This observation was further validated by high intra and interrater reliability scores as well as high correlation marks between expert ratings and automatic fluency evaluations. Consequently, the study showed that that fluency could be predicted with a high degree of accuracy.

Speech fluency was also assessed by Cucchiarini and Strik (1999). In contrast with Towell et al. (1996), whose study was examining fluency improvements in L2 speakers of French, this study takes a look at the automatic assessment of fluency in spontaneous speech of L2 speakers of Dutch. In addition to investigating the feasibility and reliability of such assessment it determines the extent to which this type of evaluation differs from the assessment of fluency in L2 read speech. The study investigated spontaneous speech of 60 non-native speakers of Dutch.²⁸ Speech samples (containing answers to eight items selected from a Dutch language proficiency test) were scored for fluency by experienced raters.²⁹ Once scored, the samples were subjected to automatic assessment by the speech recognizer according to the following criteria: speech rate, phonation/time ratio, articulation rate, total duration of sentence-internal pauses (pauses longer than or equal to 0.2 seconds), average length of pauses, number of silent pauses, and mean length of runs (average number of phones occurring between unfilled pauses not shorter than 0.2 seconds). The scores assigned by both human raters and the recognizer were then compiled and

²⁷ The number of filled pauses and disfluencies was shown to be extremely low, which was not surprising considering that the experiment dealt with read speech and that these phenomena are known to occur rarely in reading speech.

²⁸ Speakers were grouped according to their proficiency levels: lower proficiency/basic users and higher proficiency/independent users. All speakers varied in respect to sex and their L1.

²⁹ The raters were spared any specific instruction on how to assess the speakers and could listen to the samples as often as they wanted. The fluency of each sample was scored on a scale from 1 to 10.

compared with the outcomes of Cucchiarini et al. (1998). The results suggested that while automatic assessment of spontaneous L2 speech was overall a reliable indicator of perceived speech fluency (Cucchiarini et al. 1998), not all variables that were used to measure read L2 speech were effective in measuring spontaneous L2 speech. Moreover, while correlations between machine scores and human ratings were positive for spontaneous speech, their values were noticeably lower than in the case of reading speech. The authors explain this discrepancy by pointing out that the raters judging spontaneous speech had to deal with differences in grammar and vocabulary not only typically absent from read speech but also known to affect fluency ratings. In comparison with fluency variables for L2 read speech, rate of speech and phonation/time ratio in spontaneous speech are almost cut in half. The value of average length of pauses almost triples. Interestingly enough, there is hardly any change in articulation rate. In like manner, correlations between human fluency ratings and automatic fluency measures for spontaneous speech are very different for different variables. While the rate of speech, phonation/time ratio, and mean length of runs all exhibit relatively high correlations in regards to human ratings, articulation rate and average length of pauses show hardly any relation. In summary, the study shows that rate of speech, mean length of runs, and phonation/time ratio appear to be reliable indicators of fluency in spontaneous speech. It is worth noting that these three indicators express a relationship between speech and silence – a connection that is very likely at the core of utterance oral fluency.

The above study looked at the relationship between spontaneous and read speech fluency. Cucchiarini et al. (2000) examines a possible link between qualitative and quantitative fluency evaluation in L2 read speech. The authors base their inquiry on an observation that the identification of quantitative correlates of perceived fluency is key in developing objective testing instruments for fluency assessment. The study examined 20 native and 60 non-native speakers of Dutch.³⁰ The majority of non-native speakers had intermediate to advanced proficiency. Each speaker was asked to read two different sets of

³⁰ All non-native subjects lived in the Netherlands and were attending or had attended courses in Dutch. The group varied in their L1, their L2 proficiency level, age and gender. The sample was dominated by women.

five phonetically rich sentences. The recorded samples were then scored for fluency by nine experts (three phoneticians and six speech therapists). The samples were also analyzed by an automatic speech recognizer according to the following criteria: rate of speech, phonation/time ratio, articulation rate, number of silent pauses, total duration of pauses, mean length of runs, number of filled pauses, and number of disfluencies. A very high intrarater reliability coefficient for all raters involved in the experiment indicated that all raters adopted a similar definition of fluency. Due to their infrequent use, filled pauses and disfluencies were excluded from the final analysis. Out of the included markers, speech rate was proven to be the best predictor of read fluency as it correlated with expert ratings at the rate of 0.90-0.93. The other two important determinants of fluency are articulation rate and the number of pauses made. They are both related to speech rate, as articulation rate is basically speech rate without the pauses. On the other hand, when it comes to perceived fluency the frequency of pauses appears to be more relevant than their length. In other words, the difference between fluent and non-fluent speakers lies in the number of pauses they make rather than how long these pauses are. The experiment thus demonstrated that experts' ratings of reading fluency are reliable and can be predicted on the basis of quantitative measures.

The next study expatiates on the above investigation to cover spontaneous speech. In Cucchiarini et al. (2002) authors used quantitative assessment of read L2 output and spontaneous L2 speech to see how well they can predict human fluency ratings. To that end, the experiment draws on the results from Cucchiarini et al. (2000) as a base for comparison for spontaneous speech samples.³¹ Collecting the latter involved 60 non-native speakers of Dutch.³² The subjects were evenly divided in to two groups: beginners and advanced speakers. Much like in the preceding study, speakers were scored by both human raters and an automated speech recognizer. Spontaneous speech was assessed with the *Profieltoets* test comprised of questions for which the subjects had to elicit unprepared answers. Each group was given

³¹ For the purpose of this comparison only the non-native speaker data from Cucchiarini et al. (2000) was used.

³² 28 subjects were beginner and 29 intermediate Dutch L2 speakers. Three of the collected samples were of too poor a quality to assessed.

eight questions carefully selected to prompt sufficiently long answers. The beginners group was given shorter tasks, for which they were allowed 15 seconds to answer each question, taking an average of 70 seconds to complete the assessment. The intermediate group had to handle more complex questions, but its participants were given 30 seconds to answer them. On average, the subjects in this group talked for 180 seconds. Thus collected samples were then examined for fluency by 10 instructors of Dutch as a second language - five instructors per group. The instructors were informed whether they are working with beginner or intermediate material. Although the raters varied in strictness, the interrater reliability remained high. Similarly to the previous experiment, the samples were analyzed using the same quantitative fluency measures: articulation rate, rate of speech, phonation/time ratio, mean length of runs, mean length of silent pauses, average duration of silent pauses per minute, number of silent pauses per minute, number of filled pauses per minute, and number of disfluencies per minute. In comparison with reading speech, the following observations about spontaneous speech could be made: rate of speech, phonation/time ratio, and mean length of runs were all cut in half. The number of silent pauses per minute almost doubled, while mean length of silent pauses and duration of silent pauses per minute were more than doubled. Articulation rate hardly changed, disfluencies occurred a lot more frequently, as did filled pauses.³³ Strikingly, the values of the above variables for the beginners group indicate higher fluency than those of the more proficient speakers. This finding might result out of a greater cognitive demand imposed by the more difficult questions asked to the intermediate group. More cognitively demanding tasks lead to a lower speech rate, which results in lower articulation rate, shorter runs, longer pauses, etc. (Grosjean, 1980, 42-43). For spontaneous speech, rate of speech, phonation/time ratio, mean length of runs, number of silent pauses per minute, duration of silent pauses per minute all were all shown to exhibit statistically significant correlations with raters' assessment. Articulation rate and mean length of silent pauses however appear to have almost no relation to perceived fluency. The authors clarify that

³³ The frequency of filled pauses was more than 30 times higher in spontaneous than it was in read speech.

when the number of pauses in speech increases - as it does in spontaneous speech - the importance of articulation rate diminishes to the point where it can become negligible. Although still noticeable, its effects are overwhelmed by the numerous pauses committed by the speaker. In spontaneous speech the variables that explain the greatest amount of variance for the beginners were the rate of speech and phonation/time ratio, while for the intermediate subjects it's the mean length of runs. Rate of speech turned out to be the best predictor of fluency for the beginners group. For the intermediate group it was the mean length of runs that showed the strongest correlation with fluency ratings. All correlations between the automated measures and expert ratings were much lower for spontaneous speech than they were for read speech. Consistent with the previous study, the experiment demonstrated that objective automated ratings could be used to predict human ratings. However, the accuracy of such predictions was shown to be higher for read speech than it was for spontaneous oral production.

A similar investigation of linguistic variables suggestive of L2 fluency perception was conducted by Kormos and Denes in 2004. What distinguishes this study from the one summarized above is that in addition to temporal measures used in fluency evaluation it also looks at a handful of linguistic variables. L2 speech was assessed here by both native and non-native judges.³⁴ Speech samples of 16 EFL students were collected and analyzed. The subjects were divided into two groups based on their L2 proficiency (advanced and low-intermediate), each group counting eight speakers.³⁵ All participants spoke Hungarian as their first language. Students were asked to tell a story based on a cartoon comprised of between 6 and 10 pictures arranged in a logical order. Each student was given two minutes to plan their speech. Although students were allowed to talk for as long as they felt necessary, only the first two to three

³⁴ There were three native-speakers of Hungarian, all women with at least ten years of experience. In addition, there were also three native-speakers of English, including one male speaker from England, one from Scotland, and one female American.

³⁵ The advanced group consisted of eight females in their 3rd-4th year in the program. They were between 19 and 30 years old and had minimum of five years of EFL training before the experiment, 6-12 months of which was spent in an English-speaking country. They all scored high on the proficiency exam. The low-intermediate group counted six females and two males in the same age bracket as the advanced group. They all learned English in secondary school and have never been to an English-speaking country.

minutes of each verbal contribution was taped. Thus prepared recordings were then evaluated by the judges on a scale from one to five. The collected samples were also examined using the following 10 temporal and linguistic variables: speech rate, articulation rate, phonation/time ratio, mean length of runs (defined here as an average number of syllables produced in utterances between pauses of 0.25 seconds and above), number of silent pauses per minute (pauses of 0.2 seconds or longer), mean length of pauses (total length of pauses longer than 0.2 seconds divided by the total number of pauses longer than 0.2 seconds), number of filled pauses per minute, number of disfluencies per minute (e.g. repetitions, restarts, and repairs), pace (number of stressed words per minute), and pace (described as the proportion of stressed words to the total number of words). From these 10 variables only speech rate, phonation-time ratio, mean length of runs, mean length of pauses, and pace were deemed reliable predictors of fluency scores, and consequently they were the only ones investigated. Out of these, speech rate, mean length of runs, and pace were shown to be the best L2 oral fluency predictors. Mean length of pauses, although noticeably less impactful, was also statistically related to fluency scores. The analysis of the data also demonstrated that pace, speech rate, phonation-time ratio, mean length of runs, and the length of pauses were all correlated – a finding that lends further support to the importance of these variables in L2 fluency perception. In summary, the study suggests that the number of stressed words uttered per interval is a better predictor of fluency than the number of syllables uttered during said interval. This in turn seems to point out that oral fluency is primarily a temporal and intonational phenomenon. Moreover, the findings also establish a relationship between fluency and language proficiency as judges also assessed the speech samples according to linguistic variables such as accuracy and lexical diversity.³⁶ Their evaluation suggests that accuracy plays an important role in fluency judgments, as it may override the effect of

³⁶ Raters differed in regards to how much importance they attributed to accuracy, lexical diversity, and the mean length of pauses.

temporal factors on listeners.³⁷ Finally, the study shows that pausing and disfluencies do not affect the perception of fluency.

The 2004 study by Derwing et al. looks into two aspects of fluency evaluation: the relationship between temporal fluency measures and assessment by untrained raters, and how this relationship varies across different tasks.³⁸ The study engaged 20 beginner ESL learners who have lived in Canada for less than six months.³⁹ The 13 female and seven male subjects were native Mandarin speakers between 26 and 38 years of age. The subjects were asked to perform three tasks: a picture description, a short monologue, and a dialogue. For the first task students were shown a pictorial narrative comprised of eight frames. They were given 30 seconds to become familiar with it and then asked to narrate it. For the monologue tasks, subjects were invited to talk about the happiest moments of their lives. Again, they were given 30 seconds to prepare. In the last task participants had to interview the researcher about his or her happiest moment. Speech samples lasting 30 seconds were taken from the beginning of each monologue and narrative. The initial 90 seconds of each dialogue were used. Thus prepared battery of 60 stimuli was presented to the raters at a random order. There were 28 untrained⁴⁰ and three trained raters rating comprehensibility, fluency, and accentedness. The trained raters were also asked to rate the overall “goodness of prosody”. Fluency was assessed on a scale from one to nine where one indicated extremely fluent, and nine stood for extremely disfluent. A similar evaluation on a similar one to nine scale was performed in regards to comprehensibility and accentedness. Goodness of prosody was also assessed on a one to nine scale. Fluency ratings varied across tasks and were significantly lower for picture description than for the monologue and dialogue activities (there was no significant difference between ratings of

³⁷ Accuracy appears to be positively related to temporal variables affecting fluency judgments.

³⁸ The study attempted to answer the following questions: 1. Are there differences in fluency ratings across different task types? 2. Do fluency ratings of untrained judges correlate with objective temporal measures of fluency? Are “goodness of prosody” ratings related to task type? Is there a relationship between fluency ratings and L2 speakers’ self-reported exposure to L2 input? What types of relationship exists between fluency and both comprehensibility and accentedness?

³⁹ Levels 1-3 on the Canadian Language Benchmark

⁴⁰ The group consisted of 22 female and 6 male undergraduate education majors enrolled in an ESL course, between 21 and 52 years of age.

these two). The authors conjecture that these differences might reflect the degree of freedom speakers had in choosing lexical items, morpho-lexical structures and the general content. The fact that there was no significant difference in fluency ratings between the monologue and the dialogue tasks is not surprising as the two assignments were structured around the same simple topic. There was a high correlation between comprehensibility and fluency ratings (where fluency was more strongly related to comprehensibility than to accentedness), demonstrating that listener perceptions of comprehensibility are tied more to fluency than they are to accentedness. In other words, as it is easier to attend to language not interspersed with interruptions, hesitations, and false starts, heavily accented speech is likely to be better understood than speech that is disfluent. Interestingly enough, listeners' judgment of the "goodness of prosody" did not vary across tasks. In addition to being evaluated by human raters, fluency was also quantified with the following temporal measures: number of pruned syllables per second (with self-corrections, self-repetitions, false starts, filled pauses, and asides removed), number of silent pauses (>400 ms), and mean length of runs between silent pauses. Here the human rating data echoed automatic speech measurements in that the performance on the monologue and the dialogue tasks was also much better than on the picture narratives. For both former tasks, the measure of standardized pruned syllables (a composite measure with all disfluencies removed) was a successful predictor of fluency judgments, with self-repetition standing out as the least reliable temporal indicator. The study demonstrates that rating data from even untrained listeners reflects properties inherent in speech samples and can thus be used in fluency evaluation. It also shows that fluency can be evaluated as reliably as other aspects of proficiency such as accentedness or comprehensibility.

The impact of different activities on L2 oral fluency is also investigated by Iwashita et al. (2008). The study examines changes in L2 speaking proficiency across different oral tasks, and although not focusing exclusively on oral fluency, it shows how fluency along with L2 vocabulary may have the

strongest impact on perceived speaking proficiency.⁴¹ The study consisted of five tasks⁴², each task split into five proficiency levels. There were 10 oral samples recorded per each level, or 250 samples in total. After the non-audible samples have been discarded 200 total samples were left for analysis. All subjects varied in terms of age, their L1, and the length of residence in an English-speaking country, as well as in the time spent studying English prior to the study. The assessment was conducted according to the following categories: linguistic Resources (grammatical accuracy, grammatical complexity, and vocabulary), Phonology (pronunciation, intonation, rhythm), and Fluency, which was “a single feature analyzed in multiple ways”. The variables that the authors identified as suitable in measuring fluency were: filled and unfilled pauses (of one second or longer), repairs (entailing: repetitions of exact words, syllables or phrases; replacements; reformulations; false starts; and partial repetitions of words or utterances), total pausing time, speech rate, and mean length of runs. In terms of fluency, the study demonstrated that speech rate, number of unfilled pauses, as well as total pause time show a positive relationship with L2 oral proficiency. Whereas high-level learners spoke faster with less pausing and fewer unfilled pauses, no significant differences were observed for filled pauses, repairs, and mean length of runs. The study revealed that certain measures – while fairly inconsequential when in isolation – had a greater relative effect on overall scores.⁴³ Of these vocabulary and fluency were the most impactful. In summary, the study proved that while the differences between adjacent proficiency levels were not always clear-cut, it was the macro-level categories (speech rate, vocab measures, global pronunciation measure, and the global grammatical accuracy measure) that seem to have greatest influence on proficiency scores. The discussion concluded with the observation that rather than being dependent on

⁴¹ The scholarship reviewed by the study clearly shows that vocabulary and pronunciation factors are perceived as most important at lower L2 proficiency levels with fluency and grammar factors contributing little. Contributions from grammar and fluency tend to increase following the increase in L2 proficiency levels.

⁴² There were two types of tasks: independent tasks where participants were asked to express their opinion on a certain topic presented without accompanying written or aural material, and integrated tasks where participants first listened to or read information presented in the prompt and were subsequently asked to explain, describe or recount that information. The latter activities were allowed more preparation time.

⁴³ These were grammatical accuracy, vocabulary, pronunciation, and fluency (defined by unfilled pauses, total pause time, and speech rate).

one aspect of oral performance, a combination of aspects determines the assessment of the L2 oral proficiency.

The overall picture of effects that different activity types have on L2 oral fluency, emerging out of the above review is quite clear. The next study builds on these findings insofar as it investigates the scale of L2 oral fluency improvement in different learning contexts. In their 2009 experiment, Garcia-Amaya and Lorenzo examined a group of participants in a two-month long intensive overseas immersion (IM) program in Spanish. All subjects spoke English as their first language and were not allowed to speak it while participating in the program. 25 participants (16 females and nine males) volunteered to take part in the study. Five of these participants (three men and two women) were native speakers of Spanish and all were graduate students from different regions of Spain. The remaining 20 participants were divided into four groups of five students. The first group (AH1) comprised three men and two women, aged 18 to 19, with the average of 4.8 years of pre-university and 2.6 semesters of university instruction. None of the participants in this group has studied abroad. The members of the second group (AH2) have also never studied Spanish abroad. They were a bit older than the first group (average 19 to 21 years) and studied Spanish for an average of six years prior to enrolling in university. They have all received an average of 5.6 semesters of Spanish instruction. The third group (AH3) participated in an immersion program in Northern Spain for two months. These college freshmen (average 18 to 19 years) received four hours of formal instruction and two hours of various activities every day throughout their stay. They too abided by the “no-English” rule. Students in this group averaged six years of pre-university instruction and one 1 semester of university coursework. The members of the last group (AH4) also participated in study abroad programs varied in duration and Spanish L2 proficiency levels. On average they had six years of pre-university and 5.8 semesters of university-level Spanish instruction. They studied Spanish abroad for an average of 10.2 months. The subjects were asked to answer a battery of 54 questions in individual 30 to 40 minute sessions. They were given as much uninterrupted time as needed to answer each question. The topics included holiday plans, describing family members, whether

native or non-native speakers speak better, and how to prepare a particular dish. From each interview 15 longest runs were extracted and analyzed with the following temporal and lexical measures: total number of syllables per run (filled pauses excluded), total number of seconds per run (pauses included), rate of speech in syllables per second, adjusted rate of speech⁴⁴, number of filled pauses, number of repetitions, number of syllables per repetition, number of repairs per turn, number of syllables in those repairs, the total number of words uttered, the total number of words uttered in English, and the total number of syllables in these words. The results of the study were consistent with previous experiments in that students who had studied L2 abroad exhibited more fluent speech.⁴⁵ It might be worth pointing out that while the SA group could produce longer turns, the AH3 group have shown a higher speech rate, and used fewer repetitions, repairs, and pauses. No group proved to be superior with regard to all fluency measures. Whereas the AH3 group scored highest on the rate of speech (inclusive and adjusted), the ratio of filled pauses per syllable, and percentage of syllables in repetition per total syllables, the AH4 group had best scores in regards to the number of spoken words, the total number of seconds per turn, and percentage of syllables in repair per total syllables.

All of the studies presented in this section demonstrate the positive impact of live language instruction on oral fluency in the target language. Whether it is a judiciously scaffolding series of activities, classroom instruction, or a study abroad experience, communicative language practice increases oral fluency in non-native language learners. These overall improvements, expressed in increases in speech rate, longer runs, or fewer pauses, hint at a possibility that a similar advances might be occasioned by instruction in computer-based contexts – specifically by work with ASR-based pronunciation training programs. This promising idea will be addressed later in the paper.

Besides the above suggestion, there are several conclusions immediately emerging from the above literature review. The first and most general observation is that there is a clear positive relationship

⁴⁴ The total number of syllables uttered in a segment, excluding the syllables in repeats as well as in the repaired segment, divided by the number of seconds required to produce that segment.

⁴⁵ As measured by increased speech rate, higher number of words used, fewer filled pauses, repetitions, and repairs.

between overall L2 oral proficiency and oral fluency. In other words, highly proficient L2 speakers are likely to display high levels of L2 oral fluency, and conversely, fluent L2 speakers are also likely to be orally proficient. Even though they are demonstrably related, it is important to keep in mind that oral proficiency and oral fluency are two distinct, non-synchronous notions and need to be addressed as separate phenomena. This dichotomous approach also informs the design of the reviewed ASR-based CAPT software. Yet, even though all of the evaluated applications promise to improve learners' L2 oral proficiency, none of them explicitly targets oral fluency. Consequently, while there is a growing body of evidence supporting the effectiveness of these programs in the coaching of generally defined L2 oral production, little is known about how effective they are in oral fluency training.

The current study attempts to address this gap by turning to *NativeAccent* – an L2 pronunciation coaching platform listing speech fluency training among its main objectives. Featuring learning modules designed exclusively with oral fluency in mind, *NativeAccent* provides the tools necessary to investigate the extent to which computers can facilitate fluency improvement in L2 learners. Segalowitz's understanding of oral fluency as an utterance phenomenon betrays its quantifiable nature. What follows is that given the right benchmarks, one should be able to quantify improvements in learners' oral fluency. This in turn should allow one to objectively assess the effectiveness of the training program said learners have been exposed to. Straightforward though it may seem, there is one major problem with this reasoning: there is still no consensus as to which quantifiable variables are universal enough to be successfully applied across future fluency studies and dependable enough to yield reliable results (Segalowitz, 2010). The studies reviewed in the previous sections contain a handful of temporal variables demonstrated to be reliable measures of oral L2 fluency (TABLE I). Out of these, speech rate (the total number of syllables produced in a speech sample divided by the amount of time required to produce the said sample) and the number of unfilled pauses appear consistently across all studies. Furthermore, both these measures show high positive correlation with perceived oral fluency. Factoring in the relative ease

with which they can be computed, speech rate and the number of unfilled pauses seem like oral fluency benchmarks most optimal for the current study.

In order to evaluate the impact of *NativeAccent* on fluency development, the study will assess spontaneous speech samples of two groups of students: the experimental group exposed to *NativeAccent* and the control group working without the software. Speech samples recorded during the initial diagnostic and the final assessment exams will be evaluated for speech rate measured in syllables per second and the number of silent pauses per minute. Whereas both values are expected to improve for both groups (with speech rate on the rise and the number of unfilled silences decreasing), the members of the experimental group should experience greater fluency improvements than their colleagues in the control group. As pointed out by the reviewed research, such differences in improvement should be attributable to the students' work with the software. In the absence of other outside factors, the experimental group should record greater improvements in oral fluency than the control group. The fact that the program's fluency module is based in reading fluency activities should have little to no bearing on the final results as reading and spontaneous speech fluency appear to be positively correlated. In other words, students with high spontaneous speech fluency also show high reading fluency and vice versa (Cucchiarini and Strik, 1999).

V. PARTICIPANTS

All subjects participating in this experiment are graduate students at the University of Illinois in Chicago (UIC). The participants are native speakers of Mandarin enrolled in ESL 401: English for International Teaching Assistants. When applying to UIC all international graduate students are required to submit their most recent TOEFL scores indicating their ESL proficiency level. The minimum scores requirement differs by college and usually resonates with the English language proficiency needed to succeed in a given program. Students whose scores are at or below these thresholds, and who were nonetheless accepted into UIC, are required to contact the ITA (International Teaching Assistant) Center to have their ESL proficiency reevaluated. The evaluation takes place via the ITAD (International Teaching Assistant Diagnostic) test administered by the ITA program (ITAP) at the beginning of each semester. The test is graded by the ITAP staff using descriptive, qualitative criteria. The final results are kept on file and are not shared externally. Based on their performance on the ITAD, students are either cleared to fully participate in their respective graduate programs, or required to take the ESL 401 course offered each semester by the ITAP.

The selection of native speakers of Mandarin for this study was partially dictated by high enrollment numbers of Chinese students at UIC. These high enrollment numbers have also been reflected in the population of foreign students enrolled in the ITAP. Among them, the NS of Mandarin have comprised a reliably high percentage of program participants – a number that has stayed consistent over the past couple years. These steadfast enrollment numbers informed the longitudinal design of the study.

Mandarin features “flat” or undifferentiated intonation at the discourse level. As a consequence, native speakers of that language are likely to exhibit a range of other issues related to prosody and general pronunciation. In their book *Learner English: A Teacher's Guide to Interference and Other Problems*, (2001) Smith and Swan investigate some of the most common challenges faced by Mandarin-speaking learners of English. One of the problems they point out deals with reduced syllables, that - although far

less frequent in Chinese - undergo fewer phonetic changes and are pronounced more prominently by the Chinese native speakers. Consequently, Chinese ESL learners tend to stress too many English syllables, and give the weak syllables full, rather than reduced, pronunciation. Furthermore, when those same students attempt to consciously reduce the accent on the English weak forms, they oftentimes find these forms so difficult to pronounce that they avoid them altogether. Another hallmark of correct pronunciation is proper pitch intonation. Much like every tonal language, Mandarin employs pitch changes to distinguish between phonemes that otherwise would sound the same. As a result, sentence intonation varies little in Mandarin. This of course stands in stark contrast with English, where intonation patterns are employed to modify the meaning of whole utterances. Noticing and understanding the nuances of these differences has proven consistently challenging for Chinese students. Not fully understanding their arbitrariness, learners oftentimes find these intonation patterns funny and attempt to adjust them as they see fit. Their compensatory efforts, while creative, are part of the reason why their speech might sound flat or jerky to native English speakers. Lastly, because basic Chinese units of speech are monosyllabic, native speakers of Chinese tend to separate English words rather than smoothly merging them into a stream of speech. This tendency contributes to the staccato effect of the Chinese accent in ESL speech – a feature that needs a lot of attention and practice.

All subjects in this study were divided into two groups: one group (the control) enrolled in the Fall 2013 section of ESL 401 (Group 1), and the other group (the experimental) enrolled in the Fall 2014 section of the said course (Group 2). All students enrolled in the Fall 2014 section received comprehensive training on how to use the software, right at the beginning of the semester. Each group counted six participants, four females and two males, aged between 20 and 30 years old. The selection process opened with identifying all Chinese NS enrolled in the two semesters of ESL 401 and requesting their participation in the study. All of the students who agreed were then visited by the researcher and asked to sign the research consent form. The final group of participants was selected based on the two criteria: the amount of time they spent with the software, and their incoming TOEFL/ETS (English

language proficiency exams) scores. In terms of the first criterion, students who over the course of the semester spent less than one hour practicing with the program's Intelligent Tutor were not considered for the study. As far as students' performance on the TOEFL/ETS only their speaking and listening scores were taken into consideration. Generally speaking, students whose scores fell between 20 and 30 on both skills were eligible to participate in the study. Choosing that score range was consulted with the ITAP instructors and motivated by the consensus that 20-30 is most representative of their student population. As shown in TABLE II, all students in the control group met these criteria. However, when these same selection standards were applied to the experimental group many of the students failed to meet them. As evidenced in TABLE III, two of the students in the experimental group should not have been there. Nonetheless, these students were eligible to participate in the study based on their ITAD performance. As it turned out, the problem with their TESOL/ETS scores was not that they were low, but rather that they were out of date. As they were no longer representative of these students' ESL proficiency level, a more adequate form of assessment needed to be consulted. Their ITAD evaluations demonstrated that those two students' speaking and listening skills were either in line with or slightly below the 20-30 TOEFL/ETS range, thus confirming their eligibility to participate in the study. For the sake of consistency of data presentation, these students' original TOEFL/ETS scores were displayed in TABLE III. Another reason for reporting the two students' TOEFL/ETS rather than their ITAD scores, and for not sharing any information captured with the ITAD in this study, is dictated by privacy concerns. Due to their nature, including the actual ITAD evaluations – though informative - would have compromised the anonymity of the subjects. Consequently, subjects' ITAD evaluations were left out of this report. The subjects' scores were then averaged. The average scores for speaking and listening for Group 1 were 22.8 and 25.8 respectively, whereas for Group 2 they were 16.125 and 18.875 respectively (Table II and Table III). Lastly, it should be mentioned that all participants were subject to uniform curricular demands. The only difference between the two groups was that Group 2 was required to spend additional practice time with *NativeAccent v.3*.

TABLE II
CONTROL GROUP

Gender	TOEFL/ETS Speaking Score	TOEFL/ETS Listening	Academic Concentration
Female	24	25	Computer Science
Female	22	29	Computer Science
Female	22	21	Nursing
Male	22	27	Computer Science
Male	waived	waived	Biology
Female	24	27	Economics
	Average: 22.8	Average: 25.8	

TABLE III
EXPERIMENTAL GROUP

Gender	TOEFL/ETS Speaking Score	TOEFL/ETS Listening	Academic Concentration
Female	6.5	8.5	Nursing
Female	waived	waived	Mathematics, Statistics, and Computer Science
Female	23	21	Biology
Male	waived	waived	Biology
Male	15	22	Earth and Environmental Science
Female	20	24	Pharmacy
	Average: 16.1	Average: 18.9	

VI. MATERIALS

A. Classroom Materials

1. Course Curriculum

Each Fall section of ESL 401 consists of three core components. The first one comprises weekly two-hour long class sessions focusing on communication and teaching strategies necessary for effective classroom management. During these sessions students receive practice, feedback, and advice on presenting information in their field of study, as well as on adjusting sound, rhythm and overall intelligibility of their speech. The sessions also focus on developing language compensation skills that are useful when interacting with students. The second component encompasses weekly one-hour long presentation sessions that students are required to attend throughout the semester. Each student is expected to join one presentation session scheduled outside of the regular class time. During each session, students complete a 10-minute presentation. Each presentation is video-recorded for further assessment and evaluation. Upon completing their presentations, students receive feedback from both their peers and the instructor. Students are expected to incorporate this feedback into future presentation sessions and sound-work assignments. The last component encompasses independent sound work assigned by the instructor each week and to be completed via BlackBoard®. Sound-work activities are designed to improve students' general intelligibility and use of language. They give students the opportunity to focus on specific language-production difficulties and to receive individualized feedback from the instructor. These activities consist of two parts: an initial recording, and a second recording that ought to draw on the feedback given by the instructor.

In order to receive a satisfactory grade in the course students must meet a set of requirements. They have to attend at least 75% of both scheduled classes and presentation sessions. In addition, they need to complete at least 75% of independent sound-work assignments. In order to be considered completed, each of these assignments must include two

parts: an initial recording as assigned by the instructor, and a re-recording based on the individualized feedback given for the original recording. Students must also receive at least 75% on both their homework and class participation in addition to attending at least three presentation sessions. Lastly, all students have to complete the final performance evaluation administered at the end of the semester.

As there are no exams, the final grade for this course – satisfactory (S) or unsatisfactory (U) - depends on the extent of participation and “satisfactory” work students complete throughout the semester. By the end of each Fall semester course, successful students will have achieved a handful of objectives. They will have practiced teaching and language compensation strategies, culminating in conducting individual instructional presentations. They will have evaluated and analyzed communication and information exchanged from both the perspective of the audience and as a student instructor. They will have defined terms, presented field information, lead and participated in discussions, and supported explanations with examples and analogies. They will have asked and fielded questions as well as received and effectively used peer and instructor feedback. Finally, they will have made valuable connections with other students, native-speaker colleagues, and their fellow classmates.

2. Course Textbook

According to Canale and Swain (1980) communicative competence consists of grammatical competence, sociolinguistic competence, and strategic competence employed by speakers to “respond to genuine communicative needs in realistic second language situations” (p. 27). While many programs designed to prepare ITAs for work in an undergraduate classroom juggle listening, speaking, pronunciation, culture, and a myriad of pedagogical strategies, only a handful of them seem to focus on improving their students’ communicative competence (Ibid.). *English Communication for International Teaching Assistants* (2013) (ECITA) was designed to

address this shortcoming. Written specifically for ITAs working with U.S. undergraduate students, the book offers to improve ITAs communicative competence by asserting that “the task of teaching content and managing undergraduate classrooms places extraordinary demands on ITA’s L2 communicative competence (Gorsuch et al., 2011, p. 2). Yet unlike other similar textbooks (e.g. see Smith et al., 1992), ECITA attempts to boost its users’ communicative competence by focusing on the area of discourse intonation (DI). The authors defend this emphasis by pointing out that “discourse intonation plays a key role in how an instructor structures presented information and builds rapport with students” (Gorsuch, et al., 2011). They further state that discourse intonation “prioritizes intonation as the way in which we organize speech.” (p. 6). It could be thus deduced that proper discourse intonation can not only assist ITAs in effectively articulating and conveying ideas to their students, but it also makes these ideas more comprehensible – and by extension also more accessible – to the undergraduate mind. Furthermore, the authors claim that DI functions as the common denominator for language, culture, and pedagogy – the main components of classroom communication (p. 5). In other words, it “takes into account using a specific language in the context of a situation (pedagogy) directed to a specific audience (culture).” (p. 162). With such a wide-reaching impact that discourse intonation has on classroom communication it is hardly surprising that the authors decided to place it at the heart of the textbook and as a point of departure for fostering ITAs’ communication skills.

Written by academicians with extensive experience working with ITAs, ECITA is deeply rooted in Brazil’s theory of discourse intonation. A renowned British phonologist, David Brazil formulated his model of discourse intonation in effort to explain how intonation is used to convey meaning on the discourse level in the English language. His model is based on the contention that communicative value of intonation derives from the assumptions participants introduce into the interaction (Brazil, 1997). For Pickering this principle means that “in any given situation, the

worlds of the speaker and hearer will intersect to differing degrees. The extent of shared background, or *common ground*, assumed between speakers may be initially unknown, as in the case of two strangers who strike up a conversation, or may be considerable, as in the case of a discussion between family members.” (Pickering, 2001). As the tone choice indicates the extent of common ground between speakers at any particular point in the conversation, it is not hard to imagine how significant it is in those kinds of discourse where a lot of information quickly changes status from non-shared to shared – such as academic instruction. The textbook does a commendable job exploring these intonational nuances through the lens of Brazil’s theory (1997). It breaks down DI into three distinct tonal categories: falling tones (indicating a speaker’s assumption that the matter of the tone unit is a new assertion), rising tones (signifying that the speaker assumes that the matter is part of the shared background knowledge between the participants), and level tones (presenting the matter of the tone unit as neither shared, nor new, but simply as a language specimen).

The first section of the book (Chapters 1 - 5) introduces basic knowledge of DI for classroom communication. The key aspects of DI such as thought groups, prominence, pitch movement, tone, pitch level and key choice are presented in authentic context. In addition, each aspect is carefully defined and accompanied by clear rules regarding its correct use. Thanks to this thoughtful addition, students can get a better feel of how the use of these aspects impacts comprehension and mutual understanding between students and the instructor. To further illustrate the difference between the correct and incorrect use of these aspects, the book lists a variety of authentic examples recorded during undergraduate instruction. These are generally followed by creative listening and production activities, many of which are supplemented with graphic renderings of given speech phenomena captured through computer programs (pp. 24-27).

All practice materials included in the first section of the textbook are designed to help students recognize and correctly use the introduced DI components. Their design echoes

Nation's (2007) four strands of language instruction corresponding with listening practice (meaning-focused input), controlled practice (meaning-focused output), rehearsed practice (language-focused learning), and free practice (fluency development). Each chapter in this section also briefly discusses specific challenges ITAs might experience. Not only are these problems addressed according to where they are most likely to occur, but their impact on the conversation is also briefly evaluated. Lastly, all chapters in this section conclude with test presentation assignments including a self-evaluation rubric. This form of self-assessment seems particularly helpful as it accounts for pronunciation, sociolinguistic, textual, and functional competence features – collectively helping students develop the full range of communicative competence.

While the first section of the book provides a very comprehensive summary of the basic aspects of discourse intonation, a couple of its chapters could use additional explanations. For instance, the book associates rising intonation with shared knowledge. Though this may indeed be one way of marking shared knowledge, the authors seem to ignore the relationship between intonation and the types of questions instructors may want to ask their students. For example, Celce-Murcia et al. (2010) point out that questions starting with *who*, *what*, *where*, *when*, or *how* are usually asked with falling intonation. Rather than discussing this important detail, the authors have ITAs complete an activity structured as an intonation game. Although the game only features WH- questions, students are asked to employ rising intonation “whenever it may be appropriate” (p. 50). The chapter on intonation also does not discuss what happens when shared (old, given, distressed, or presupposed) information is positioned *after* the prominent word. According to Bolinger (1986, p. 126-127), such words should be pronounced in a flat manner. Rather than explaining this rule, the book tunes out the words occurring after the prominence. Finally, when presenting the aspect of prominence itself the authors cite the following rule, “In a new topic, the last meaningful word is usually prominent” (p. 25). Although this is correct,

readers are left wondering what makes a word “meaningful”. This overlooking might have been easily avoided by pointing out that meaning in English is mainly encoded in content words such as verbs, nouns, adjectives, and adverbs. While these issues are significant, they could be easily fixed in class or as a part of homework.

The second section (Chapters 6 - 9) of the book shows ITAs how to effectively use discourse intonation tools covered in the first section in authentic teaching situations. It does that by confronting students with typical teaching tasks such as: “Introducing Yourself and Your Course”, “Leading Labs and Classes”, “Giving Instructions and Advice”, and “Asking and Answering Questions” (p. v – vi). Each chapter in this section demonstrates how different aspects of discourse intonation could be used to effectively perform a given task. Similarly to the first section, each chapter also contains a handful of ingenious activities (e.g. outlining and organizing presentations, utilizing charts and figures, signaling and phrasing questions, and giving examples) aimed at helping ITAs internalize the aspects of DI in relation to classroom instruction. Those activities are also sequenced from listening and analysis, through controlled and rehearsed practice, all the way to free practice, which in this case translates to mock mini lessons or presentations. All activities in this section end with an ITA performance assessment and a metacognitive self-reflection exercise.

The third and last section of the book (Chapters 10-12) includes additional resources for both the ITAs and their instructors. Chapter 10 introduces strategies useful in resolving communication breakdowns. As opposed to the first two sections, the strategies presented in this chapter do not need to be practiced and are meant for immediate use. Considering that communication breakdowns pose one of the biggest challenges to ITAs (Gorsuch et al., 2010) this chapter analyzes the most common reasons for why communication falls apart. These include problems with pronunciation, vocabulary, fluency, grammar, and giving answers to unanticipated questions. Aside from furnishing practical strategies for solving communication breakdowns, the

chapter also clarifies that speech fluency is not as much about speed as it is about “speaking at a rate that is appropriate for the audience and for the material one is trying to teach” (p. 162).

Chapter 11 spells out techniques to improve overall discourse intonation, showing ITAs how to continue improving their skills beyond the course. The two techniques laid out in this chapter are transcribing and mirroring (where students are asked to mimic a recording of a native speaker in terms of the speech patterns as well as the body language). Thoughtful use of these two techniques ought to help students develop useful metacognitive habits as they work their way up the academic hierarchy. Finally, the last chapter of the book is devoted to assessing ITA performance.

Designed by a group of experienced university instructors, this this 178-page long course book has several other outstanding features. First of all, unlike some other textbooks with sparse exercise sections (see Smith et al., 1992), ECITA contains a wealth of thoughtful activities designed to help ITAs practice the newly acquired knowledge. Such practice-oriented design gives ITAs plenty of opportunities to put the theory into action.

Secondly, the text emphasizes metacognition as an important learning and teaching tool. The authors define metacognition as related to thinking about, planning, and managing one’s own teaching process (Gorsuch et al., 2010). The included metacognitive exercises teach ways of self-evaluation, self-improvement, peer-evaluation, and meaningful methods of using instructor feedback. In addition, they foster effective instruction by teaching how to use outlines, transcripts, microteachings, recordings, and mirroring. Each chapter also concludes with a self-reflection activity inviting learners to think about the most difficult aspects of a specific discourse intonation component as well as the best way to surmount those difficulties.

Thirdly, one of the biggest selling points of this textbook is the attention it pays to the cultural dimension of a typical American college classroom. Already in the first chapter, the authors make students consider their own expectations as well as issues they expect to encounter

in their classrooms (p. 3). Such inclusion of cultural expectations should improve ITAs' teaching skills as well as prepare them to address any cultural misunderstandings in a polite and culturally sensitive fashion.

Lastly, the textbook comes with a DVD containing both audio and video material. All samples have been recorded in real classroom and show native as well as non-native English speakers teaching a wide range of subjects. It is worth reiterating that rather than using native speakers as paragons of correct language use, the material also employs non-native speakers to illustrate appropriate pronunciation. Viewing fellow non-native English speakers as examples of correct speech motivates students by showing them a very real, attainable goal. The course materials also take advantage of technology other than the provided DVD by utilizing websites such as "YouTube" or "Americanrhetoric.com". For instance, the chapter on speech paragraphs (Chapter 5) asks students to watch a speech by George Clooney accessible via "Americanrhetoric.com". Such reliance on Internet sources not only makes the course materials more relevant to student life, but it also makes students more reflective of their future teaching responsibilities. Students might also appreciate additional teaching tips and advice scattered across the book.

It is also worth noting here that the material is based on authentic ITA reports informed by student concerns ranging from correct pronunciation of American last names, through how to understand undergraduate slang, all the way to handling unanticipated questions. Thanks to these personal touches students might feel that their needs are thoughtfully addressed. Overall, the book seems to deliver on its promise of addressing communicative competence through meaningful discourse intonation training. Thus, it provides a valuable resource for both ITAs and their instructors.

B. Online Materials

One of the biggest challenges witnessed by modern language instruction is that of the finite nature of resources available to instructors. In particular, time seems to be of the essence. With the length of instruction limited by institutional constraints, contact time with the target language becomes a prized commodity, dedicated to the types of activities that are most effective when performed in class. Unsurprisingly, inductive grammar instruction, or communicative language practice take here precedence over reading or writing – both of which could be practiced outside of classroom with relatively high effectiveness. Perhaps the one type of instruction that is most commonly sidelined, is that of correct pronunciation. Curbed by time constraints, language instructors tend to marginalize that aspect of oral production, leaving it up to the students to practice it in their own time. Regrettably, majority of students lack the resources to follow through on that additional practice, leaving that aspect of their language education largely ignored. By giving students access to *NativeAccent v.3* we were hoping to address that shortcoming. The idea behind the implementation of *NativeAccent v.3* was to provide the ITAP students with an easily accessible set of tools they could use to improve their ESL pronunciation. Though designed independently from the textbook, the software focuses on a handful of pronunciation areas (e.g. lists, correct pausing, or time reduction techniques) that are also the focal points of the textbook, while also highlighting other areas crucial to native-like, English pronunciation. To that end, *NativeAccent v.3* is a useful tool, adequately complimenting the program spelled out in the course textbook.

1. NativeAccent v.3

NativeAccent v.3 is an online language-training platform focusing on ESL pronunciation coaching. It uses both ASR and pronunciation feedback modules, which makes it as a fully functional CAPT suite. The program concentrates on four major curricular areas: pronunciation, grammar, word stress, and fluency.

Before starting their work with *NativeAccent*, students are required to take the program's diagnostic test aimed at identifying their ESL proficiency level.⁴⁶ The test measures 38 phonemic elements, 28 grammar skills, seven fluency skills and word stress. It takes about an hour to finish and should be completed within one sitting. Upon the completion of the initial assessment, students are presented with a "skills report" broken down into the said four curricular areas. The results are used to generate a lesson plan specifically tailored to each participant's unique speech pattern. The lesson plan is created by the *Intelligent Tutor* - an algorithm analyzing speech samples from the point of view of the four curricular areas - prioritizing the areas/skills in greatest need of improvement. Thus created curriculum encompasses 35 hours of on-line training distributed evenly over the course of 13 weeks and amounting to two and a half hours of practice per week. Each week of training contains 10 lessons that students are asked to complete at their own pace and convenience. Instruction is delivered in 15-minute lesson bites, each lesson emphasizing one of the four main curricular areas.

Student performance on each lesson is evaluated according to the following color scheme: green = good, yellow = close, red = in need of improvement. Successful completion of a given lesson unlocks the next assignment. If their oral production is lacking, students are encouraged to follow the pronunciation suggestions offered by the program. During each lesson students can monitor their performance by comparing their own recordings to the native speaker samples conveniently linked on the main taskbar. Completing the program's self-assessment is a star system used to both indicate student progress and to visually evaluate their skills: one star = less than 25% correct, two stars = less than 50% correct, three stars =

⁴⁶ Considering how students' TOEFL scores were not always the best barometer of their language proficiency, and how privacy concerns precluded the ITAD scores from being used in this study, it might be tempting to defer to *NativeAccent*'s internal diagnostic for a more reliable quantitative baseline. While this is a noteworthy idea, such test needed to have been administered to both groups of students, which was impossible in the context of the current study.

less than 80% correct, and four stars = over 80% correct. Upon the completion of all lessons in the sequence, students are asked to take the final assessment, which is identical to the program's initial diagnostic exam.

The current study concentrates on fluency, an aspect of oral proficiency, which *NativeAccent* addresses with three categories of activities, focusing on linkage, pausing and fluent production. The first category – linkage – refers to the melding of words when speaking fluently. For example, the collocation "did you" when uttered by a native speaker is often pronounced as "did jew". This phenomenon is referred to as word linkage or word liaison. Word melding is governed by a number of standard patterns, and the purpose of this category of lessons is both to teach students to be able to recognize fluent speech containing targeted words, and to produce it. In other words, rather than getting students to say "Djeet?" for "Did you eat?" the program teaches them to not pause after every word, and when appropriate, to run some of the words together. The lessons in this category include:⁴⁷

a.) Linkage - Consonants and Vowels

When normal conversation occurs, the phonemes in a sentence sound as though they form one, long word. Only when there is a clear pause can word boundaries be recognized. Conversational speech flows smoothly and connectedly. To be able to produce such fluent speech, students need to learn how to connect sounds and reduce the length of vowels when appropriate. For instance, when a consonant at the end of a word is followed by a vowel at the beginning of the next word, the consonant gets slightly reduced and sounds as though it were the first sound in the next word. Students should also recognize that speaking too precisely gives a stilted, unnatural sound to speech.

⁴⁷ The detailed information regarding oral fluency lessons comes from an interview with Gary Pelton, the CTO of Carnegie Speech, conducted in November 2014.

b.) Linkage – Reduction

Native English speakers tend to shorten words that are less important in fluent speech. These words usually include pronouns, articles, certain verbs like *to be* and auxiliary verbs (can, will, should, etc.), as well as conjunctions. They get shortened by turning the core vowel into a “schwa” /ə/ sound and pronouncing the word faster.

c.) Linkage – Palatalization

A negative contraction followed by the personal pronoun *you* blends the contraction's final “t” /t/ with the pronoun's initial “y” /j/ creating the “ch” /tʃ/ sound. Due to its diminished importance, the personal pronoun *you* is reduced in the process. For instance, “didn't you” will sound like “didn chew”, and “couldn't you” will be pronounced like “couldn chew”. When the final consonant of the word before the personal pronoun *you* is a voiced consonant, the pronoun is pronounced as though it were spelled with a “j” /dʒ/. For example, “did you” will sound like “di jew” and “should you” will be uttered as “shou jew”.

d.) Linkage - H-Reduction

This is a special case that occurs often enough to merit its own lesson type. It is not uncommon to omit the “h” /h/ sound in fluent speech for certain word combinations. For example, “Isn't he?” is pronounced as “Isn tee?”, while “Wouldn't he?” usually turns into “Wouldn tee?”.

The second category – pausing – is quite self-explanatory. Non-native speakers often do not pause at appropriate places, which might cause their speech to be hard to understand. The

lessons in this group are designed to give students pointed practice in appropriate pausing.

a.) Pausing in sentences

Pausing in conversation is a natural way to take breaths while speaking and to add a layer of prosodic meaning to the conveyed message. The most natural place to take a breath is between sentences, clauses, or items on a list. The easiest way to remember when to pause is to think of it in terms of punctuation. If a conversation were written down, one could see where the punctuation would demarcate sentences and clauses. This is also where pauses ought to be taken. Research shows that pauses in fluent speech fall between 0.2 seconds and 2 seconds.⁴⁸

b.) Pausing in lists

Pauses also occur between items on a list. When enumerating items on a list the pitch should rise slightly on the first syllable of each item and fall on the last.

The last category – fluent production – consists of two lesson types: focusing on speech rate and concentrating on time reduction. The goal of these lessons is to make students aware of their disfluencies. The version of *NativeAccent* employed in the current study was only able to recognize long pauses and speech rate. According to the program’s developers, future versions of the software should also be able to detect “disfluencies such as repeated and missed words” and broadly understood “poor pronunciation”.

⁴⁸ For more on pausing in American English speech, consult Nakatani et al. (1981).

a.) Speech rate

The rate of speech in regular conversation is known to affect fluency perception. If a speaker is too slow, his or her speech will sound labored and will be difficult to follow. Conversely, if the rate of speech is too fast, the listener will have a hard time processing the content, let alone understanding the speaker. An average rate of speech benefits everyone; clear pronunciation and natural rhythm facilitate speech production, while connected thought groups separated by pauses aid comprehension. Average speech for adult native speakers of American English may vary depending on many factors but according to the National Center for Voice and Speech it hovers around 150 words per minute.⁴⁹

b.) Time reduction

Being able to speak concisely and fluently in a non-native language takes a lot of practice and skill. Paul Nation's 4-3-2 exercises are one of the few techniques that have been shown to be effective in getting ESL students to produce more fluent speech (Nation, 2007). *NativeAccent v.3* implements activities informed by Paul Nation's findings. Exercises in this section provide an opportunity to practice all the fluency skills introduced to students throughout the course, including the appropriate rate of speech, pausing, linking, and putting correct stress on words. In each exercise, students are asked to communicate the same information three times in response to the same question. Each time they are given a shorter time period in which to respond. The goal is for students to improve the organization and fluency of their contributions each time they repeat the information.

⁴⁹ Source: <http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/quality.html>

VII. PROCEDURE

A. *NativeAccent* – Scoring and Analysis

On top of following the course's regular curriculum, students in the experimental group were asked to practice with *NativeAccent*. Practice with the software was neither mandatory nor was it supervised by the ITAP faculty. Instead, student performance was automatically recorded and evaluated by the program. Thus compiled data was then used to assess students' overall performance in the course. Out of the eight fluency markers employed by *NativeAccent*, pausing in sentences and speech were chosen to quantify oral fluency improvement.

Regarding pausing in sentences, *NativeAccent v.3* relies on the built in speech recognizer designed to segment the audio input in 10 ms frames. For any silence shorter than 100 ms, the program assumes that there is no pause. For any silence lasting longer than 500 ms, the algorithm concludes that the pause must be due to "some other factor" and thus marks it in red as not belonging in fluent discourse. Consequently, all pauses between 100 and 500 ms are considered a reasonable attempt to pause in a sentence and thus marked in either green (meaning "good") or yellow (meaning "close"). It should be mentioned that the software also detects when users have made pauses longer than two seconds in their speech. The sections of sound spectrum contained between each two pauses are recognized as words. They constitute the basis for measuring speech rate, which the program assesses in words per minute (wpm). As far as this measure is concerned, values falling between 140 and 180 wpm are considered fluent speech and marked in green. This value range is also consistent with the average speech rate of adult NS of American English suggested by the National Center for Voice and Speech.⁵⁰ Any values falling outside of this range, above or below, are marked as red. Since users are generally able to understand how fluency readings change based on their input, the program assumes that providing numeric values is enough to indicate how close a learner is to the desired threshold.

⁵⁰ Source: <http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/quality.html>

Improvements in students' fluency were tracked and recorded by the program throughout the semester. In addition, the software recorded the time spent on the electronic assessments including the initial diagnostic exam as well as on any other consecutive tests students chose to complete. The total time each student spent with the program was also measured. The difference between the latter and the total time spent solely on the electronic assessments was recorded and interpreted as the total time students spent on practicing with the *Intelligent Tutor*. Fluency values were computed in terms of the probability of correct pronunciation and expressed as percentage values. They were arranged according to the eight fluency focal areas and listed as individual scores. A distinction was made between the scores yielded by electronic assessments and those resulting from semester-long practice with *Intelligent Tutor*. The *Intelligent Tutor* scores were referred to as "current values". Their arithmetic average across eight focal categories generated the "current fluency score" indicative of the participants' end-of-semester fluency level. This value was then compared with the averaged fluency score from *NativeAccent's* initial diagnostic exam. The difference between these two scores reflected the participants' progress over the course of the semester. All values were generated based on sound samples produced by students to complete each activity. The number of such samples was also recorded.

The curriculum furnished by *NativeAccent* concluded with a final assessment administered 14 weeks after students had begun their work with the program. The exam was identical to the initial diagnostic test used by the program's *Intelligent Tutor* to evaluate users' ESL proficiency and generate personalized curricula. A comparison between the final scores and diagnostic results served to illustrate students' overall improvements. Indirectly, it also testified to the program's effectiveness.

B. ITAD and the Final Presentations - Scoring and Analysis

1. ITAD

Students' improvement in oral fluency was also measured based on the difference between their performance on the ITAD exam and that on their final presentations. The two assessments were administered to all students enrolled in ESL 401.

The ITAD exam is administered to newly admitted graduate students whose English language proficiency might prevent them from fully realizing their potential in an academic setting. The goal of the test is to diagnose students' oral production skills as well as their general intelligibility. The test has been developed by the faculty managing the ITAP. It is administered at the beginning of each fall semester and conducted according the procedure outlined in APPENDIX A. The ITAP faculty is also responsible for determining the format of the test and for keeping the exam up-to-date. Teaching courses and mentoring students also belong to ITAP faculty's responsibilities. All students participating in the current study underwent the ITAD examination during the second week of their respective semesters. The exam consists of eight, free-response questions, distributed to students in a separate booklet. A copy of the booklet is attached in APPENDIX B. Students are allowed to familiarize themselves with the questions prior to answering them. In interest of time, the exam is administered in groups, within which students are expected to work individually, record their own answers, and refrain from consulting with others. Students are also not allowed to use the Internet. Once they are familiar with the assignment, students are given exactly one minute to answer each question. Since the test is administered online, all answers are digitally recorded. The software employed to this end is Java-based *Wimba Voice Board* accessible via the ITAP's *BlackBoard* platform. As for the hardware, students recorded their answers using stereo headsets. Thus recorded answers are evaluated by the instructors supervising the test. Each answer is assessed based on a number of criteria corresponding to its theme. For instance, the first question asks for the best time of the

year to visit the examinee's hometown. In order to adequately address this inquiry during the limited time given, students have to be able to organize their ideas, express those ideas using the "should/could" construction, do so in a suggestive tone, and employ appropriate vocabulary relating to seasons and holidays. The more of these criteria are featured in an answer, the higher the score. A list of focal points assessed in each question is included in APPENDIX C. Instructors are expected to record their evaluations on standardized rating sheets distributed prior to the study. The evaluation sheets are listed in APPENDIX D. The exam's final results are then shared with individual students via email or over the phone. Since ITAD is an internal exam, the results are not shared with any other parties, and are discarded after one semester.

Student performance on the ITAD was recorded and analyzed for speech rate expressed in syllables per second and the total number of silent pauses per minute. Participants were given one minute to respond to each one of the eight questions. The analysis of their replies produced the value of speech rate and the number of silent pauses per reply. These two values were then averaged for each participant and applied to the exam as a whole.

2. Final Presentations

Final presentations took place 12 weeks after the ITAD exam. To pass, students needed to conduct an oral presentation on a topic related their field of expertise. The topics were pre-assigned so that students were given a sufficient amount of time to adequately prepare their presentations. Each student was given 10 minutes to present. Presentations were conducted in English. They were all videotaped and subsequently evaluated against the ITAD test results from the beginning of the semester. For each recorded presentation, the video and audio tracks were carefully separated using iMovie.⁵¹ Since only the opening segment of each recording featured

⁵¹ The program and its documentation can be downloaded from:
<https://itunes.apple.com/us/app/imovie/id408981434?mt=12>.

uninterrupted speech, the first minute of each presentation was isolated for further analysis. Thus obtained speech samples were twice filtered for background noise using *Audacity* – a free audio-editing program available for both Mac and PC platforms.⁵² All sound samples were saved at 128 kbps and 44100 Hz. They were all one-minute long.

3. Scoring and Analysis

Recordings were evaluated with *Praat* – an open-source speech-analyzing program.⁵³ Both speech rate and the number of pauses were measured automatically using the “Speech Rate” script written in order to automatically assess speech rate in a wide-ranging longitudinal study performed at the University of Amsterdam.⁵⁴ The said study demonstrated that the accuracy of *Praat*’s automatic assessment is high enough for the software to be used as a reliable speech analysis tool.⁵⁵ It also partially explains why human raters were not used in this study. Another reason against employing human raters goes back to Segalowitz’s definition of utterance fluency anchoring this study. Engaging human raters would expand that definition to include perceived fluency, thus exceeding the scope of this study. Finally, as demonstrated by Cucchiarini and Strik (1999), raters judging spontaneous speech don’t just focus on speech fluency but oftentimes find their evaluation affected by grammatical and lexical features of the assessed sound samples.

To gauge speech rate without the need to transcribe the sound sample, *Praat* automatically detects syllable nuclei. To that end it first scans the sample for all sound peaks above a threshold defined automatically based on the median intensity (measured in decibels (dB)) of that sample. These peaks are then marked as potential syllable nuclei. Since this sort of

⁵² The program and its documentation can be accessed at: <http://audacity.sourceforge.net>.

⁵³ The program and its documentation are available at: <http://www.fon.hum.uva.nl/praat/>.

⁵⁴ The updated version of the script used in the current study is available at: <https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2>

⁵⁵ For more on Praat, see: De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2), 385-390.

labeling often results in multiple peaks within each syllable, the script rejects the peaks that are not demarcated by dips of 3 dB in intensity.⁵⁶ False syllable nuclei may also be detected in place of loud fricatives. To prevent this mislabeling the script automatically discards all non-voiced peaks. The value of -25 dB was chosen as the silence threshold, as it is not only consistent with previous research (De Jong and Wempe, 2009) but it also provides most accurate readings considering the amount of background noise in the analyzed samples. Finally, 0.3 s was chosen as the minimum pause length as it falls between 0.2 s and 0.4 s – the two values adopted by other studies in the field. It also falls right between 0.1 s and 0.5 s, which is consistent with the threshold used by *NativeAccent v.3* in evaluating speech samples for silent pauses. Thus configured, Praat was able to reliably determine the number of syllables and silent pauses in each sound sample. Speech rate was calculated by dividing the former by the total length of each sample in seconds and expressed in syllables per second. Silent pausing was expressed as a ratio of the number of pauses over the total sample duration in seconds.

It is important to mention here that only those students who met all study eligibility criteria had their speech samples submitted for analysis. The total number of 24 samples was analyzed. Those included: six ITAD samples from the control group, six ITAD samples from the experimental group, six final presentation samples from the control group, and six final presentation samples from the experimental group. The analyzed samples, though varying in quality were suitable for analysis. No sample was rejected on the basis of its quality.

⁵⁶ Repeated trial and error established the value of 3 dB to provide most accurate account of the number of peaks per syllable and syllables per word.

VIII. RESULTS

The data produced by this study was compiled and analyzed using several statistical tests. All statistics for this study were performed with *Microsoft Excel*, and *SPSS*. Descriptive statistics were run for both the control and experimental groups. The data for each group corresponded to ITAD and Final Assessment scores for the speech rate (expressed in syllables per second) as well as the number of silent pauses (expressed in the number of silent pauses per minute). The descriptive statistics comprising the mean and the standard deviation for the control group are listed in TABLE IV. Corresponding values for the Experimental Group can be found in TABLE V.

As displayed in TABLE IV, students in the Control Group made on average 27.27 silent pauses/minute during their initial assessment. By the end of the semester, that number increased to 33.17 silent pauses/minute, indicating that students were pausing more at the end of the course than at its onset. This in turn suggests a decrease in oral fluency. An opposite trend could be observed in terms of speech rate. Here students started the semester with an average of 3.18 syllables/second, and ended it with 3.42 syllables/second, demonstrating clear fluency gains. The extent to which these two fluency indicators offset each other exceeds the scope of this project. The question of which one of these two benchmarks is the *better* measure of oral fluency, should also be considered in future research.

The results for the Experimental Group, shown in Table V, were far less ambiguous. On average, the subjects in this group made 45.58 silent pauses/minute at the beginning of the instruction. That number dropped over the course of the semester to 44 silent pauses/minute, pointing to fluency gains. This fluency improvement was echoed by gains in speech rate. On average, students in this group began the semester with the speech rate of 2.72 syllables/second, which increased to 2.95 syllables/second by the semester's end. The fact that by the end of the semester, students in this group were speaking faster while making fewer pauses clearly demonstrates that their oral fluency improved.

TABLE IV
DESCRIPTIVE STATISTICS FOR THE CONTROL GROUP

	Silent Pauses		Speech Rate	
	Initial Assessment (ITAD)	Final Assessment	Initial Assessment (ITAD)	Final Assessment
Mean	27.27	33.17	3.18	3.42
Standard Deviation	13.33	11.13	0.53	0.21

* N = 6 across all groups

TABLE V
DESCRIPTIVE STATISTICS FOR THE EXPERIMENTAL GROUP

	Silent Pauses		Speech Rate	
	Initial Assessment (ITAD)	Final Assessment	Initial Assessment (ITAD)	Final Assessment
Mean	45.58	44	2.72	2.95
Standard Deviation	21.09	17.96	0.71	0.38

- N = 6 across all groups

The goal of the current study was to examine the effectiveness of *NativeAccent v.3* on the ESL oral fluency in adult native speakers of Mandarin. In order to answer the research question which sought to determine whether students' oral L2 fluency improved as a result of their work with *NativeAccent v.3* three groups of non-parametric tests were performed. Due to the small, unequally distributed sample other tests - such as the t-test - were deemed unsuitable for the analysis.

In order to assess whether the two groups improved over the course of the semester the Wilcoxon signed-rank test was conducted within each group. The null hypothesis assumed that participants' performance has not changed between the diagnostic ITAD exam conducted at the beginning of the

semester and the final assessment administered at the end of the semester. The tests were run at the alpha level of $\alpha = 0.05$. The tests performed for the control group showed no statistically significant change in the mean values of speech rate and silent pauses, indicating that students in this group did not improve on their oral fluency. Similarly, the tests performed for the experimental group also showed no statistically significant change in said values, and consequently no oral fluency improvement. The outcomes of these tests are summarized in TABLE VI.

TABLE VI

CHANGE IN FLUENCY VALUES WITHIN GROUPS (INITIAL VS. FINAL ASSESSMENT)

Compared Groups	Test Performed	Significance	Decision
Control Group: Silences	Wilcoxon signed-rank test	0.249	Fail to reject the null
Control Group: Speech Rate	Wilcoxon signed-rank test	0.249	Fail to reject the null
Experimental Group: Silences	Wilcoxon signed-rank test	0.917	Fail to reject the null
Experimental Group: Speech Rate	Wilcoxon signed-rank test	0.463	Fail to reject the null

* N = 6 across all groups

With the Wilcoxon signed-rank test demonstrating no statistically significant changes within the examined groups, the next step was to measure the difference between the groups' performance on the final assessment. To that end the groups were compared using the Independent-Samples Mann-Whitney U test. Four tests were completed: one comparing the difference in the number of silences using the initial ITAD score, one comparing the difference in the number of silences using the final assessment score, one evaluating the variance in the value of speech rate using the initial ITAD score, and the other one evaluating the variance in the value of speech rate using the final assessment score. The only statistically significant difference detected was between the mean final assessment scores for speech rate. At a

significance level of $\alpha = 0.05$ the reported significance was 0.041. Because 0.041 is less than 0.05, it is safe to say that the null hypothesis can be rejected and to conclude that there is a statistically significant difference between these two groups (TABLE VII).

TABLE VII

CHANGE IN FLUENCY VALUES BETWEEN GROUPS (INITIAL AND FINAL ASSESSMENT)

Compared Groups	Test Performed	Significance	Decision
Control vs. Experimental – Silences - ITAD	Mann-Whitney U Test	0.065	Fail to reject the null
Control vs. Experimental – Speech Rate - ITAD	Mann-Whitney U Test	0.310	Fail to reject the null
Control vs. Experimental – Silences – Final Assessment	Mann-Whitney U Test	0.394	Fail to reject the null
Control vs. Experimental – Speech Rate – Final Assessment	Mann-Whitney U Test	0.041	Reject the null

* N = 6 across all groups

Finally, consideration was given to assessing the relationship between the total time spent with *NativeAccent* (TABLE VIII) and fluency improvements spelled out by changes in participants' speech rate and the number of silent pauses as per the final assessment (TABLE IV and TABLE V).

TABLE VIII
TIME SPENT WITH NATIVE ACCENT V.3

Student	Total Time (Active Learning Time)	Time Spent on Assessments	Time Spent with Intelligent Tutor
Student 1	1453 min	57 min 28 sec	1395 min 32 sec
Student 2	411 min	23 min 54 sec	387 min 6 sec
Student 3	405 min	29 min 16 sec	375 min 44 sec
Student 4	489 min	38 min 16 sec	450 min 44 sec
Student 5	182 min	35 min 19 sec	146 min 41 sec
Student 6	777 min	23 min 15 sec	753 min 45 sec
Mean	619 min 30 sec	34 min 35 sec	584 min 55 sec
Standard Deviation	411 min 47 sec	11 min 47 sec	403 min 51 sec

The study attempted to examine whether the time spent with *NativeAccent v.3* had an impact on or in any other way contributed to fluency improvement reflected in increased speech rate or reduced number of pauses. To address this question, nonparametric correlations were conducted using Spearman's rho (TABLE IX).

TABLE IX
THE EFFECT OF THE TIME SPENT WITH NATIVEACCENT V.3 ON ORAL FLUENCY

	Speech Rate		Silences	
	r_s	Significance	r_s	Significance
Total Time Spent Learning	0.771	0.072	-0.086	0.872
Time Spent on Assessments	0.029	0.957	0.812	0.050
Time Spent on Intelligent Tutor	0.771	0.072	-0.086	0.872

* N = 6 across all groups

Table VIII demonstrates that there is a significant correlation between the time spent on assessments and the decreasing number of silences. While it appears that total time spent on learning and time spent on Intelligent Tutor correlate to increased speech rate, the correlation is not significant. Likewise, the time spent on assessments showed no significant correlation with the improvement in speech rate. A similar lack of significance in correlation can be observed between the change in the values of silences and the total time spent learning as well as the time spent on Intelligent Tutor.

IX. DISCUSSION

The goal of this study was to examine the effectiveness of ASR CAPT software on the oral fluency of adult native speakers of Mandarin enrolled in an ESL course at a large research university. All participants were graduate students working on advanced degrees in a variety of fields. The software employed was *NativeAccent v.3*, but the results obtained may be relevant to other similar programs. While the reviewed research demonstrates that ASR CAPT software is an effective – if still imperfect – tool in L2 oral proficiency training, none of these studies looked into how effective such software is in improving students' oral fluency. To the author's best knowledge there is currently no research investigating this matter. Consequently, it is difficult to relate the results of the current inquiry to any previous findings.

The current study employed two quantitative measures in assessing its participants' oral fluency: speech rate (syllables per second) and silent pauses (number of pauses per minute). The two values were chosen based on how well they correlate with perceived oral fluency as assessed by human raters. The selection was based on previous research, reviewed in SECTION IV and summarized in TABLE I. The data collected in the current investigation suggest that the members of the control group⁵⁷ on average experienced no statistically significant improvement in oral fluency as measured in speech rate and in the number of silent pauses. This however does not mean that the control group did not improve at all. In fact, TABLE X indicates that the majority of this group improved their mean speech rate scores.

⁵⁷ The control group only received instruction in a face-to-face learning environment and did not have access to the software for additional practice.

TABLE X
CHANGE IN SPEECH RATE FOR THE CONTROL GROUP

Student	ITAD Mean	Final Assessment Mean	Change
Student 1	3.264	3.416	0.151
Student 2	3.762	3.55	-0.212
Student 3	2.187	3.183	0.996
Student 4	3.473	3.166	-0.307
Student 5	3.169	3.566	0.397
Student 6	3.238	3.65	0.412

The control group showed high standard deviation of speech rate for both the ITAD Mean and the Final Assessment Mean, 0.533 and 0.206 respectively. Interestingly enough, the majority of the control group also registered an increase in the number of silences (TABLE XI), an outcome that likely offsets the gains in the mean speech rate. The control group also showed high standard deviation for both the ITAD Mean and the Final Assessment Mean, 13.323 and 11.125 respectively. Since the participants could not be randomly assigned to the groups and because the groups were small the final results may have been impacted.

TABLE XI
CHANGE IN SILENCES FOR THE CONTROL GROUP

Student	ITAD Mean	Final Assessment Mean	Change
Student 1	35.75	17	-20.75
Student 2	8	27	19
Student 3	44.75	30	-14.75
Student 4	27.375	47	19.625
Student 5	20.375	44	23.625
Student 6	27.375	34	6.625

Very similar results could be observed when students had the option to practice with the software. The overall mean improvement in oral fluency for those students also proved statistically insignificant. In terms of speech rate, half of the sample registered an increase, while the other half showed a decrease in its mean values (TABLE XII). The standard deviation for the ITAD Mean and the Final Assessment Mean were found to be 0.711 and 0.382 respectively.

TABLE XII

CHANGE IN SPEECH RATE FOR THE EXPERIMENTAL GROUP

Student	ITAD Mean	Final Assessment Mean	Change
Student 1	3.446	2.7	-0.746
Student 2	2.539	3.233	0.694
Student 3	1.556	2.716	1.16
Student 4	3.103	2.483	-0.62
Student 5	2.358	3.05	0.692
Student 6	3.301	3.5	0.199

A parallel outcome could be observed in regards to the number of silences, which decreased only for half of the group. The other half of this group experienced an increase in silence values, which likely helped counteract any fluency gains (TABLE XIII). The standard deviation for the ITAD Mean and the Final Assessment Mean were found to be 21.086 and 17.956 respectively.

TABLE XIII
CHANGE IN SILENCES FOR THE EXPERIMENTAL GROUP

Student	ITAD Mean	Final Assessment Mean	Change
Student 1	29.625	75	45.375
Student 2	52.75	28	-24.75
Student 3	84.5	41	-43.5
Student 4	37.75	40	2.25
Student 5	41.25	53	11.75
Student 6	27.625	27	-0.625

Much like in the control group, the lack of statistically significant fluency improvement in the experimental group does not indicate that oral fluency did not improve for its individual members. Indeed, some members registered an overall improvement in both fluency measures. Since the participants were not able to be randomly assigned to the groups and because the groups were small the final results may have been impacted.

To summarize the above findings, let us first take a look at fluency improvement as measured in the number silences per minute. As far as this criterion is concerned, the two groups did not show a statistically significant difference. In other words, while both groups registered a change in the number of silences the change was not substantial enough to be statistically significant and thus to prove a positive effect of the time spent with the software.

What was statistically significant was the difference in speech rate improvement between the two groups. The data shows that the experimental group registered a greater improvement in speech rate than the control group. Even though the improvement was modest, it was statistically significant and thus may be linked to the positive effects of *NativeAccent v.3* on L2 oral fluency (TABLE VII).

The above observation is not reflected in the degree of correlation between the time spent with the software and the mean speech rate values (TABLE VIII). In other words, there is no evidence that the

time spent with the software translated into higher speech rate values (TABLE XIV). Consequently, if the increase in the experimental group's speech rate cannot be related to the time the members of this group spent with *NativeAccent*, the reasons for their improvement are to be sought elsewhere. Seeing how the participants in this group were overall less proficient ESL speakers as indicated by their incoming TOEFL and ITAD (diagnostic exam) scores, their speech rate improvement might be attributed to factors such as higher motivation or out-of-class interactions. It is also possible that the students in this group improved more precisely because they were starting from a lower proficiency level than their colleagues in the control group. Further, qualitative inquiry in form of student interviews might be helpful in shedding light on discrepancies in student improvement as well as the factors that influence it.

When it comes to the relationship between the time spent with *NativeAccent* and the number of silences, the study showed a strong, positive correlation. In other words, the longer students practiced with the program, the less they were pausing. This trend is particularly telling when one considers the time spent on *NativeAccent's* assessments (as opposed to the platform's Intelligent Tutor), indicating that the longer students worked with software-internal test, the less they had to pause (TABLE VIII). This positive correlation corroborates the effectiveness of the software's automated feedback where user's pausing patterns are rated using a three-color scale. Any silences registered to last longer than 500 ms are marked as red and not characteristic of fluent discourse. Unfilled pauses shorter than 500 ms but longer than 100 ms are considered reasonable attempts to pause in a sentence and consequently marked either as green ("good") or yellow ("close"). For any silences lasting shorter than 100 ms, the recognizer assumes that there is no pause and provides no feedback. This support for automatic feedback, even if unsophisticated, seems especially noteworthy as it contributes to the discussion of the significance of testing in general, and ASR-based automatic review in particular.

Encouraging though these results are, it is important to take a look at the instances where the additional work with *NativeAccent* correlated with a decrease in the participants' fluency. In terms of speech rate, Students 1 and 4 demonstrate fluency loss (TABLE XI). This observation is further

supported by a decline in fluency as expressed by an increase in the number of silent pauses. Here, in addition to Students 1 and 4, a decrease in fluency was also demonstrated by Student 5 (TABLE XII). Part of the reason for this drop in fluency might have to do with the conditions of the final assessment. It is not unreasonable to assume that for a number of the students taking the course, this was the first time they were required to conduct an academic presentation in English – a presentation that they were being evaluated on. Such stressful circumstances might have caused some of the students to speak more deliberately, to pause more frequently in order to carefully choose their words. Being in a position where one must really focus on what one sounds like, might cause speakers to forego fluency for the sake of delivering a clearer message, which might help explain fluency loss registered for Students 1 and 4.

Lastly, the effectiveness of the software in oral fluency training was confirmed by the program's built-in assessments. According to *NativeAccent's* internal tests, all students improved on their fluency scores. The average improvement was 14% (TABLE XIV). However, the fact remains that this development is not confirmed by the independent assessment performed by at the beginning (ITAD exam) and at the end of the semester (final presentations). This discrepancy might result out of the conditions addressed in the next section.

TABLE XIV

FLUENCY IMPROVEMENT ACCORDING TO NATIVEACCENT V. 3

Student	Total Time (Active Learning Time)	Time Spent on Assessments	Time Spent with Intelligent Tutor	Mean Intake Scores (Based on 34 samples)	Mean Final Scores	Average Improvement
Student 1	1453 min	57 min 28 sec	1395 min 32 sec	76%	95%	19%
Student 2	411 min	23 min 54 sec	387 min 6 sec	62%	81%	19%
Student 3	405 min	29 min 16 sec	375 min 44 sec	70%	90%	20%
Student 4	489 min	38 min 16 sec	450 min 44 sec	76%	80%	4%
Student 5	182 min	35 min 19 sec	146 min 41 sec	68%	72%	4%
Student 6	777 min	23 min 15 sec	753 min 45 sec	74%	92%	18%

X. LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

Despite its promising design the study fell somewhat short of expectations. The reasons for this are manifold and will be addressed in the following paragraphs. The discussion of the limitations of this study is by no means exhaustive and much like the study itself, it is intended as a point of departure for a larger debate on the role of ASR CAPT software in a modern foreign/second language classroom.

The current study set out to measure the impact of CAPT software on L2 oral fluency in adult learners. To that end it relied on *NativeAccent v.3*, a CAPT platform designed to coach ESL pronunciation. The training consisted of an automatically generated, individually tailored curriculum encompassing 35 hours or 2100 minutes of practice over the course of a semester. Whereas students were asked to complete the whole training, doing so was not made mandatory. As a result, none of the students considered for this study successfully completed the entire curriculum. This trend was not limited to Mandarin NS students, as the majority of the students enrolled in ESL 401 completed less than two out of the recommended 35 training hours. This casual approach to work with the software significantly limited the sample of students eligible to participate in the study to those who completed at least an hour worth of *Intelligent Tutor* assignments. One gets the sense that had work with *NativeAccent* been made a mandatory component of the curriculum, the experimental sample would have been much greater and thus more reflective of the program's impact.

Related to the shape of the practice is the matter of its frequency. As demonstrated by Ellis (2002) language students who practice in frequent, regular intervals are significantly more likely to outperform students whose practice habits are irregular, infrequent, and uneven in length and intensity. When applied to the current study, the participants who followed their individualized curricula in the prescribed fashion were bound to improve more than their counterparts who approached the task with less discipline. In particular, students who practiced irregularly, and in long bouts were likely to see their improvement lagging behind their more disciplined colleagues. Seeing how the participants likely placed between these two extremes, their heterogeneous study patterns might have affected their final

performance. To account for this limitation, future studies ought to enforce a practice framework standardized for all participants.

As work with *NativeAccent* did not yield statistically significant improvement in fluency scores of the experimental group over the control group, it is possible that other factors might have influenced the participants' progress. Among those, the time spent communicating in the second language outside of classroom might have had an effect on students' oral proficiency. Such effect might have been positive - resulting in an increase in fluency scores; or negative - causing those scores to go down. Since students' extracurricular activities were not monitored in this study, it is impossible to determine the extent to which they might have affected students' oral performance. Although there is no sensible way to control student L2 interactions outside of classroom, one possible way to account for their impact would be to ask students to record all instances of their L2 communication. Such records could be then considered a valuable source of input and included in data analysis.

Not surprisingly, the short amount of time that students devoted to work with the software showed no significant effect on the final fluency scores. Due the low number of students that followed the practice guidelines, the sample was small. Had more students attempted to follow the practice guidelines, the size of the sample would have been larger, which in turn would have made it more likely to obtain statistically significant results. Because only a small section of students completed adequate practice time, the eligible sample size was small, making the data susceptible to extreme values. Again, making the interactive component of the program mandatory would have likely increased the sample eligible for analysis and produced statistically significant results.

The two groups differed in their fluency scores on the ITAD exam, with the values for the experimental group significantly lower than the values for the control group. This difference was also echoed by the TOEFL scores, with the experimental group boasting lower numbers than the control group. Aware of their wanting performance on TOEFL and thus presumably mindful of their lacking ESL proficiency those students might have been more motivated to improve their English with all means

available. Since *NativeAccent* provided an additional opportunity to advance their language skills, those students with low incoming scores might have been more motivated to use the program in the first place and consequently they might have ended up logging more practice time than other participants. This relationship between students' motivation and the time spent with the software might help explain the lower incoming scores of the experimental group. To address this shortcoming, future studies should ensure that the participants' intake scores are similar in distribution so as to provide a better point of comparison for analysis. Randomization is also another way to address this issue. Had the control and experimental groups been randomized, the groups would have been more representative of the entire sample population. Unfortunately, randomization was not feasible due to the nature of this study. In regards to motivation, future studies might investigate how user-friendly and motivating students found the materials. This could include a qualitative or a mixed-method approach. It is sensible to assume that students who did not find the software particularly intuitive, user-friendly, or exciting, would be less inclined to use it on a regular basis. This in turn might have negative impact on their progress and the resulting scores.

Another limitation rising from small sample size is the potential impact of the participants' field of expertise on their oral performance. It is safe to assume that students specializing in subjects requiring significant and regular oral production might be more skilled in conducting presentations or addressing questions requiring free-response answers. A quick glance at the participants' concentrations reveals that none of them are majoring in humanities, communication studies, or other disciplines demanding intensive oral production. It is possible that expanding the sample to include participants studying such subjects would have an effect on the experiment' final outcome.

All students participating in the study were native speakers of Mandarin. A tonal language unrelated to English, Mandarin poses its own unique challenges to the learners of English of a second language. Especially when it comes to mastering oral fluency, English proves particularly challenging for Mandarin NS, as it relies on different suprasegmental, sentence-level, and discourse-level intonation

patterns than English. Seeing how Chinese students would have to learn those patterns before improving their fluency, including NSs of other languages in the sample might have produced different final outcomes, and deliver a more representative set of data.

The initial oral fluency values were obtained from the ITAD exam – an eight-question test, where students were given one minute to answer each question. The corresponding final values were calculated based on final presentations that students had to give at the end of the semester. The presentations lasted between 20 and 30 minutes and were all videotaped. Since only the opening of each presentation featured a speech sample suitable for analysis, the first minute of each lecture was evaluated for the purpose of the study. The two tests were thus very different instruments – a difference, which might have affected the experiment's results. Future experiments should benefit from having the same test serving as pre and post-assessment.

The quality of sound samples recorded for the study was not always immediately suitable for computer analysis. Although the quality of all recordings was sufficiently high for evaluation by human raters, all sound samples had to be filtered for background noise. In addition to digital noise filtering, recordings of the final presentations had to be isolated from video files - a process that may have had an impact on their quality. Whereas all final presentation recordings featured one student at a time, the ITAD samples were recorded in large groups, where all students were oftentimes responding to questions at the same time. Naturally, such recording conditions substantially contributed to the background noise and in turn had an impact on the quality of individual ITAD recordings. Future experiments can easily address this shortcoming by recording students individually and in isolation.

Another factor that might have impacted the quality of the recordings was the quality of the hardware that the samples were registered with. All ITAD files were recorded using PCs, equipped with USB microphones and headsets. The platform used for recording was *BlackBoard Wimba Voiceboard* run via the *Firefox* Internet browser under Windows 7. All samples were registered using the same parameters (128 kbps at 44100 Hz) and encoded as .mp3 files. Whereas all hardware and software used

in the process was homologous, some headsets were older or more worn down than others, which might have caused sound quality differences among the recorded samples. As the final presentations were registered with a digital video-recorder, the sound quality of these samples was also different from that of the ITAD exam recordings. Future research can address this incoherence in sound quality by ensuring that future studies use the same equipment to record all samples.

The final limitation has to do with *Praat* - the free, open-source software used in audio analysis. “Speech Rate” – a script written for *Praat* to analyze speech rate and the number of silences was employed to evaluate the recordings for fluency. Even though research has demonstrated *Praat*'s effectiveness in sound sample analysis, it also pointed out a margin of error that although statistically insignificant with larger samples, might have an effect on samples smaller in size. The authors of the script also draw attention to the fact that the results are most accurate if the analyzed samples are clear and without background noise. Apart from ensuring the high quality of the recorded samples, future studies might confront this shortcoming by having the values of speech rate and the number of pauses computed by hand by experienced human raters.

XI. CONCLUSION

The goal of this study was to examine the effectiveness of ASR CAPT software in oral fluency training. The software employed for this purpose was *NativeAccent v.3* designed as a pronunciation-training platform for the speakers of English as a second language. The study evaluated two groups of native speakers of Mandarin enrolled in various graduate programs at a large research university in the United States. All subjects were enrolled in the two consecutive Fall sections of ESL 401 – a course designed for international teaching assistants (ITAs) and emphasizing oral language production and pronunciation skills. The control group was selected from among the Fall 2013 cohort and had no access to the software. The experimental group was part of the Fall 2014 cohort and had considerable exposure to the software. Both groups followed the same curriculum, belonged to the same age bracket, and spoke the same first language. The thesis started with an assessment of the overall effectiveness of ASR CAPT software in ESL training. In addition to demonstrating the software's efficacy, the survey pointed out that while CAPT technologies do generally improve L2 pronunciation, little is known about their impact on L2 oral fluency. The study attempted to address this research gap by answering the following research question: How effective is CAPT software (*NativeAccent v. 3*) in improving oral L2 fluency?

Although the study showed no statistically significant improvement overall, a handful of students did improve after all. And while a few subjects did experience modest fluency improvements, others saw their fluency scores drop. Notably, the experimental group registered an improvement in speech rate over the control group, however as this development cannot be related to the time spent with *NativeAccent* its cause is best left for future investigation. Though *NativeAccent* did register an average improvement in its users' fluency values, the boost was not reflected in fluency assessments conducted independently at the beginning and at the end of the semester.

The study had a considerable number of limitations that curtailed the eligible research sample. Since the goal of this study was to assess the efficacy of ASR CAPT software, it was assumed that its participants would spend an adequate amount of time working with the program. Regrettably, this was

not the case. Records show that students only practiced for an average of 10 hours out of the prescribed 35 hours of training per semester. As a result, only those students who spent more than one hour interacting with the program were considered for the study. This limitation, combined with the students' reluctance to consenting to have their data used in the study, led to the total sample counting only 12 students. While there was little else that could have been done to boost students' participation, making practice with *NativeAccent v.3* a mandatory component of the curriculum would have certainly resulted in a greater sample, more representative data, and more conclusive results. Distributing 35 hours over the course of 14 weeks yields 2.5 hours of practice a week or about 20 minutes a day – a very reasonable time investment given the potential benefits. Still, even with such a limited sample and against an array of limitations, the study did show *NativeAccent's* positive – if very modest – effect on L2 oral fluency. As the program was not used extensively, these modest results are a promising invitation for future research into the matter.

In terms of its possible pedagogical implications, ASR CAPT software such as *NativeAccent* holds a great promise for language instruction. The reviewed studies have shown that when intelligently incorporated into regular instruction, such software can be a great time-saver. As teaching pronunciation to L2 learners tends to be a time-consuming enterprise, CAPT programs have the potential to relieve instructors of that duty, allowing them to concentrate on more urgent and consequential objectives. Although specialized software will not solve all the problems plaguing modern language classroom, it does promise to bring a much-needed relief to overextended language curricula.

CITED LITERATURE

- Anderson, J. R. (1976). *Language, memory, and thought*. Hillsdale, NJ: Lawrence Erlbaum.
- Bernsen, N. O., Hansen, T. K., Kiilerich, S., & Madsen, T. K. (2006). Field Evaluation of a Single-Word Pronunciation Training System. *The Fifth International Conference on Language Resources and Evaluation, LREC 2006* (pp. 2068-2073).
- Bolinger, D. (1986). *Intonation and its parts: Melody in spoken English*. Stanford, CA: Stanford University Press.
- Brazil, D. (1997). *The communicative value of intonation in English*. Cambridge: Cambridge University Press.
- Brown, H. D. (2007). Chapter 2:. In *Teaching by principles: An interactive approach to language pedagogy* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall Regents.
- Canale, M., & Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics, 1*(1), 1-47.
- Celce-Murcia, M., Brinton, D., & Goodwin, J. M. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. Cambridge: Cambridge University Press.
- Cucchiari, C., & Strik, H. (1999). Automatic assessment of second language learners' fluency. *Proceedings of the 14th International Congress of Phonetic Sciences*. San Francisco, 1-7 August 1999. 759-762.
- Cucchiari, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America, 107*(2), 989-999.
- Cucchiari, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America, 111*(6), 2862-2873.

CITED LITERATURE (CONTINUED)

- Cucchiaroni, C., Strik, H., Boves, L. (1998) "Qualitative Assessment of Second Language Learners's Fluency: An Automatic Approach", in *ICSLP 98, Proceedings of the 5th International Conference on Spoken Language Processing*. Sydney Convention Centre, Sydney, Australia. 30th November - 4th December 1998. CD Rom edition. Rundle Mall: Casual Production. Paper, 752. Vol. 6, 2619-2623.
- De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2), 385-390.
- De Jong, N. H., Schoonen, R., & Hulstijn, J. (2009). Fluency in L2 is related to fluency in L1. Paper presented at the Seventh International Symposium on Bilingualism, Utrecht, The Netherlands.
- Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second Language Fluency: Judgments on Different Tasks. *Language Learning*, 54(4), 655-679.
- Derwing, T. M., Thomson, R. I., & Munro, M. J. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System*, 34(2), 183-193.
- Eisler, F. G. (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. New York: Academic Press.
- Ellis, N. C. (2002). Frequency Effects in Language Processing. *Studies in Second Language Acquisition*, 24(02), 143-188.
- Fathman, A. (1980). Repetition and correction as an indication of speech planning and execution processes among second language learners. In H. W. Dechert & M. Raupach (Eds.), *Towards a cross-linguistic assessment of speech production*. Frankfurt am Main: Lang.
- Fillmore, C. J. (1979). On Fluency. In C. J. Fillmore, D. Kempler, & W. S. Wang (Authors), *Individual differences in language ability and language behavior* (pp. 85-101). New York: Academic Press.
- Fox, A. (2000). *Prosodic features and prosodic structure: The phonology of suprasegmentals*. Oxford: Oxford University Press.

CITED LITERATURE (CONTINUED)

- García-Amaya, L. (2009). New findings on fluency measures across three different learning contexts. In *Selected proceedings of the 11th Hispanic linguistics symposium* (pp. 68-80).
- Goldman-Eisler, F. (1958). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, 10(2), 96-106.
- Gorsuch, G. (2010). *English communication for international teaching assistants*. Long Grove, IL: Waveland Press.
- Grosjean, F. (1980). Temporal variables within and between languages. In H. W. Dechert & M. Raupach (Authors), *Towards a cross-linguistic assessment of speech production* (pp. 39-53). Frankfurt am Main: Lang.
- Grosjean, F. and A. Deschamps. 1972, 'Analyse des variables temporelles du français spontané'. *Phonetica*, 26. 129-156.
- Grosjean, F. and A. Deschamps, 1975. 'Analyse contrastive des variables temporelles de l'anglais et du français Vitesse de parole et variables composantes, phénomènes d'hésitation.' *Phonetica*, 31. 144-184.
- Hieke, A. E. (1985). A Componential Approach to Oral Fluency Evaluation. *The Modern Language Journal*, 69(2), 135-142.
- Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2007). Assessed Levels of Second Language Speaking Proficiency: How Distinct? *Applied Linguistics*, 29(1), 24-49.
- Kim, I. (2006). Automatic Speech Recognition: Reliability and Pedagogical Implications for Teaching Pronunciation. *Educational Technology & Society*, 9(1), 322-334.
- Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145-164.
- Kormos, J. (2006). Fluency and Automaticity in L2 Speech Production. In *Speech production and second language acquisition* (pp. 154-165). Mahwah, NJ: Lawrence Erlbaum Associates.

CITED LITERATURE (CONTINUED)

- Lennon, P. (1990). Investigating Fluency in EFL: A Quantitative Approach. *Language Learning*, 40(3), 387-417.
- Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Author), *Perspectives on fluency* (pp. 25-42). Ann Arbor: University of Michigan Press.
- Luoma, S. (2004). Speaking Scales. In *Assessing speaking* (pp. 59-95). Cambridge: Cambridge University Press.
- Nakatani, L. H., O'Connor, K. D., & Aston, C. H. (1981). Prosodic aspects of American English speech rhythm. *The Journal of the Acoustical Society of America* 69(S1).
- Nation, P. (2007). The Four Strands. *Innovation in Language Learning and Teaching*, 1(1), 2-13.
- Neri, A., Cucchiarini, C., & Strik, W. (2003). Automatic speech recognition for second language learning: how and why it actually works. In *Proc. ICPhS* (pp. 1157-1160).
- Neri, A., Mich, O., Gerosa, M., & Giuliani, D. (2008). The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, 21(5), 393-408.
- Neri, Ambra, Catia Cucchiarini, and Helmer Strik. "ASR-based Corrective Feedback on Pronunciation: Does It Really Work?" In *Interspeech 2006 - ICSLP*. Proc. of The 9th International Conference on Spoken Language Processing, Pittsburgh, PA., 2006. 1982-985. Print.
- O'Connell, D. (1980). Cross-linguistic investigation of some temporal dimensions of speech. In H. W. Dechert & M. Raupach (Eds.), *Towards a cross-linguistic assessment of speech production*. Frankfurt am Main: Lang.
- Pica, T. (1994). Questions from the Language Classroom: Research Perspectives. *TESOL Quarterly*, 28(1), 49-79.
- Pickering, L. (2001). The Role of Tone Choice in Improving ITA Communication in the Classroom. *TESOL Quarterly*, 35(2), 233-255.

CITED LITERATURE (CONTINUED)

- Precoda, K., Halverson, C. A., & Franco, H. (2000). Effects of Speech Recognition-based Pronunciation Feedback on Second-Language Pronunciation Ability. (pp. 102-105).
- Raupach, M. (1987). Procedural Knowledge in Advanced Learners of a Foreign Language (R. Towell, Ed.). In J. A. Coleman (Ed.), *The advanced language learner: Papers of the Joint AFLS (Association for French Language Studies)/SUFLRA Conference held at the Roehampton Inst., London in April 1986*. London: Center for Information on Language Teaching and Research.
- Rehbein, J. (1987). On fluency in second language speech. In H. W. Dechert & M. Raupach (Authors), *Psycholinguistic models of production* (pp. 97-105). Norwood, NJ: Ablex Pub.
- Riggenbach, H., & Koponen, M. (2000). Overview: Varying perspectives on fluency. In *Perspectives on fluency* (pp. 5-24). Ann Arbor: University of Michigan Press.
- Sajavaara, K. (1987). Second language speech production: Factors affecting fluency. In H. W. Dechert & M. Raupach (Authors), *Psycholinguistic models of production* (pp. 45-65). Norwood, NJ: Ablex Pub.
- Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition*, 14(4), 357-385.
- Segalowitz, N. (2010). Measuring L2 Oral Fluency. In *Cognitive bases of second language fluency* (pp. 29-52). New York: Routledge.
- Skehan, P. (2003). Task-based instruction. *Language Teaching*, 36(1), 1-14.
- Smith, B., & Swan, M. (2001). Chinese Speakers. In *Learner English: A teacher's guide to interference and other problems* (pp. 310-324). Cambridge; New York: Cambridge University Press.
- Smith, J., Meyers, C. M., & Burkhalter, A. J. (1992). *Communicate: Strategies for international teaching assistants*. Englewood Cliffs, NJ: Regents/Prentice Hall.

CITED LITERATURE (CONTINUED)

- Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-273). Philadelphia: John Benjamins Publishing.
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17, 84-119.
- Ullman, M. (2015). The Declarative/Procedural Model: A Neurobiologically-Motivated Theory of First and Second Language. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction*. New York, NY: Routledge.

APPENDICES

APPENDIX A

ITAD Admin Instructions

I. Set Up

1. Set up before students come- about 15 minutes prior to their arrival
2. Create a folder in dropbox- name the folder the day and time of ITAD, add department if necessary *Ex: July 25 3:00PM ITAD Chemistry*
3. Create a folder for each student
4. Within each student folder, copy and paste Clarity Recorder

II. Once Students Arrive- Test Practice and Preparation

1. Have students take a seat at every other computer
2. Advise students not to use the Web
3. Hand out cover sheet for ITAD and have students fill it out with their basic information- make sure they identify whether or not they have a TAship.
4. Walk through file-saving process- have students practice as a group
 - a. Go into student folder
 - b. Click on clarity recorder
 - c. record your name and department
 - d. save in desktop under July 25/your folder/# of question for file name "0"
 - e. Press "new" button on the Clarity Recorder to begin recording a new sound clip
5. Have students practice saving another question on their own
 - a. Go into student folder
 - b. Click on sound recorder
 - c. Record the answer to "What are you going to do after you finish today?"
 - d. Save in desktop/July 25/your folder/file name "00"
 - e. Press "new" on the Clarity Recorder to begin recording a new sound clip
6. Have students check to see that they can hear themselves on the recorded practice questions
 - a. Give these tips:
 - i. Make sure that you are not too close to the microphone
 - ii. Make sure that you speak loudly
 - iii. Use your headphones to block out your neighbors

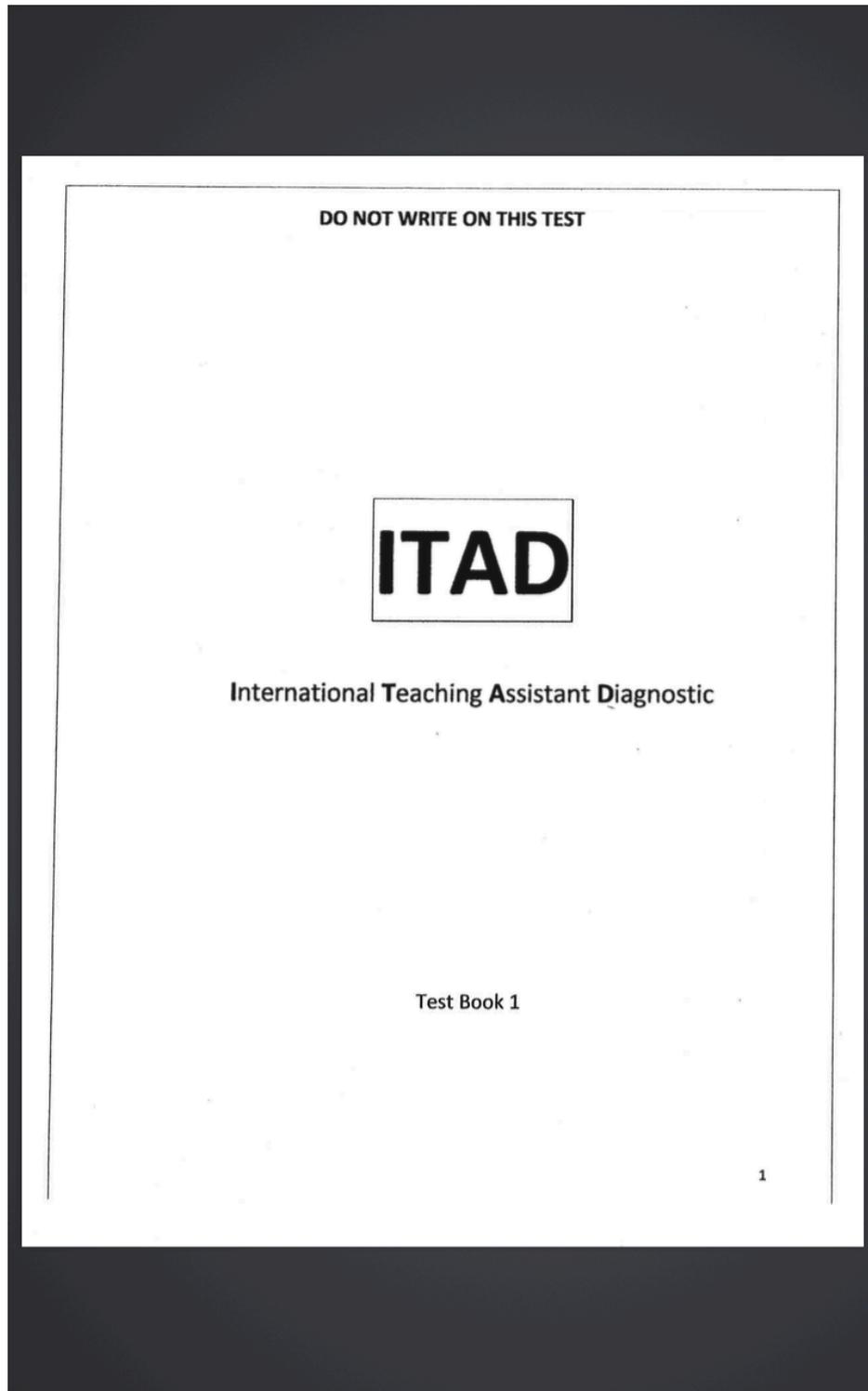
III. ITAD Test

1. Hand out ITAD test booklets and explain that we are using the test to determine general intelligibility
Ex: "Pretend you are leaving a voicemail for somebody. You are leaving important information in the voicemail and you want people to listen only once and understand the message. Speak a little more slowly, and do not worry about using perfect grammar."
2. Read the ITAD directions aloud, slowly
 - a. Ask students not to write inside their ITAD test booklets
 - b. Advise students that if they do need scratch paper in order to take notes or organize their thoughts, they may use the backside of their ITAD cover sheet.
3. Tip for the proctor- after questions 1, 2, and 3 check to make sure each student is saving their questions properly

APPENDIX A (continued)**IV. ITAD File Saving and Clean-Up**

1. Advise students that the test is over and to clean up their workspaces (put headphones back, etc.)
2. Let students know their scores will be sent to them via email or a phone call, whichever they listed as their preference on the ITAD cover sheet
3. Save all ITAD test files by moving the folder *"July 25 3:00PM ITAD Chemistry"* from the drop box on to a disc

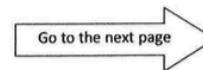
APPENDIX B



APPENDIX B (continued)**DO NOT WRITE ON THIS TEST****TEST 1**General Directions

In this test, you will be able to show how well you communicate in English. The test will last approximately twenty minutes. The test proctor will read the questions which are printed in the test booklet and you will have one minute to answer each question. These questions may not be directly related to your field, but they are designed to assess your general oral English proficiency.

Your score will be based on your recorded speech sample. Concentrate primarily on communicating your ideas clearly.



APPENDIX B (continued)

DO NOT WRITE ON THIS TEST**NOW THE TEST WILL BEGIN:**

You will hear the question one time. You will have 15 seconds to prepare your answer on a piece of paper. Then, you will have 60 seconds to answer the question. Please speak clearly and directly into the microphone.

1. I'm planning a trip to your home city, but I don't know when the best time would be to visit. Can you recommend a time of the year that would be best to visit in terms of weather or activities.
2. A younger cousin is trying to choose between attending a large university or a smaller college. What are some of the things that your cousin should consider when making this decision?

APPENDIX B (continued)

DO NOT WRITE ON THIS TEST

3. Now please look at the six pictures below. I'd like you to tell me the story that the pictures show, starting with picture number 1 and going through picture number 6. Please take one minute to look at the pictures and think about the story. Do not begin the story until I tell you to do so.



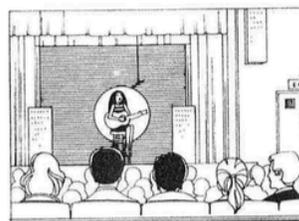
1



2



3



4



5



6

Page 4 of 6

4. What could the people have done if the concert had been cancelled?

APPENDIX B (continued)**DO NOT WRITE ON THIS TEST**

5. Explain a term from your field and give an example or an analogy to help me understand it clearly.

6. Describe a current topic or development in your field, and explain why it's important.

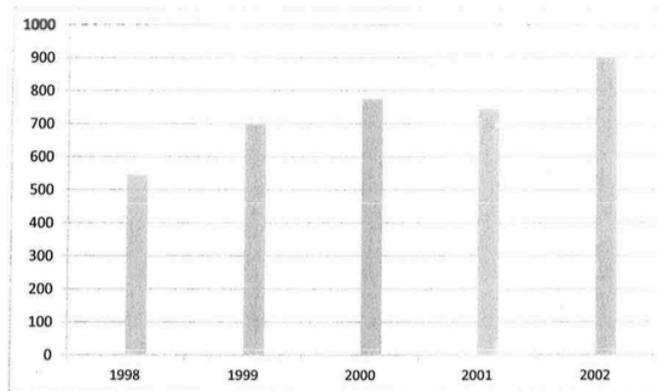
APPENDIX B (continued)

DO NOT WRITE ON THIS TEST

7. Imagine that you had scheduled a meeting with your advisor, and two hours before the scheduled time, you realized that you would not be able to meet with him/her. Call him/her to apologize and reschedule the meeting.

8. The graph below represents the number of American students studying abroad over a period of time. Tell me about the information in the chart and what trends you predict for the future.

American Students Studying Abroad



Page 6 of 6

Go to the next page

6

APPENDIX C

ITAD Elicitation Guide

1. Visiting Hometown
 - a. Organization of thoughts
 - b. “you should/could” structure
 - c. Suggestive tone
 - d. Vocabulary: season, time of year, vacation type activities
2. Large or small university
 - a. Organization of thoughts
 - b. Advice giving tone/knowledge
 - c. Presenting argument/reasons
3. Tell a story based on pictures
 - a. Consistent tense
 - b. Elaboration
 - c. Connectors – not describing each picture individually
4. If concert had been cancelled
 - a. Conditional modals/conditional past perfect – *They could have...*
 - b. Connectors/clear cohesion between content words – *and/or/also, etc*
 - c. Vocabulary: other activities, tv, movie, etc.
5. Field term
 - a. Clear presentation format
 - b. Organization of ideas
 - c. Pronunciation of key terms
 - d. Working vocabulary for on-the-spot description
6. Topic/development and its importance
 - a. Clear description of topic
 - b. Address *why* it’s important
 - c. Organization – main idea to details
7. Rescheduling meeting
 - a. Discourse tone: apologetic, respectful
 - b. Structure – *could we/can I/I would like to*
 - c. Stating reason for not being able to make it
8. Describe Chart and trends
 - a. Structures: *more/less than, increase/decrease, future tense*
 - b. Inferences based on chart
 - c. Overall trends, not details of each year

APPENDIX D

ITAD Rating Sheet

Student: _____ Rater: _____

Total score: _____ Averaged score: _____

Quest #	Score	Language / Production	General Comments
1			
2			
3			
4			

APPENDIX D (continued)

Quest #	Score	Language / Production	General Comments
5			<input type="checkbox"/> lost final consonants <input type="checkbox"/> lost final "s" <input type="checkbox"/> lost final reduced ending <input type="checkbox"/> shortened clear vowels <input type="checkbox"/> lost medial syllables <input type="checkbox"/> unclear consonant articulation <input type="checkbox"/> word stress off <input type="checkbox"/> lack of strong final stress <input type="checkbox"/> uneven timing <input type="checkbox"/> choppy – needs longer connectors <input type="checkbox"/> pace too fast <input type="checkbox"/> too few pauses <input type="checkbox"/> too many restarts <input type="checkbox"/> intonation
6			<p>LANGUAGE USE</p> <input type="checkbox"/> patterns off <input type="checkbox"/> missing function words <input type="checkbox"/> lack of verb tense continuity <input type="checkbox"/> lack of verb tense sensitivity <input type="checkbox"/> lack of elaboration <input type="checkbox"/> trouble finding words <input type="checkbox"/> too much interpreting/guessing required of listener
7			<p>STRONG SKILLS</p> <input type="checkbox"/> good general sound <input type="checkbox"/> good music <input type="checkbox"/> good overall pacing <input type="checkbox"/> full responses
8			<p><i>Circle three to focus work on.</i></p>

VITA

- NAME:** Szymon Edward Zuberek
- EDUCATION:** B.A., Germanic Studies and Linguistics, University of Illinois at Urbana-Champaign, 2009
 B.A. Political Science, University of Illinois at Urbana-Champaign, 2009
 M.A., Germanic Languages, Literatures, and Linguistics, University of Illinois at Chicago, 2015
 M.A., Applied Linguistics, University of Illinois at Chicago, 2016
- ACADEMIC EXPERIENCE:** German Language Lecturer, Foreign Language Department at North Park University, Chicago, IL, 2015
 Graduate Researcher and Administrator, Institute for the Humanities at University of Illinois at Chicago, IL, 2013-2015
 German Language Instructor, Department of Germanic Studies at University of Illinois at Chicago, IL, 2012-2013
 Graduate Researcher, Language and Culture Learning Center at University of Illinois at Chicago, IL, 2011-2012
 Undergraduate Instructor, Global Studies Initiative, College of Liberal Arts and Sciences, University of Illinois at Urbana-Champaign, IL, 2009
- CONFERENCES:** Zuberek, S. (2015) Amerika, Du Hast es Besser? – An Exploration of the Image of America in the Post-Second-World-War German Poetry, Interdisciplinary Graduate Student Conference “Converging Narratives: The Personal Meets the National”, University of Illinois at Chicago, IL.
 Zuberek, S. (2014) Adopting a Monster – An Analysis on the Role of the Uncanny in Kleist’s “Der Findling”, The 23rd annual Graduate Student Symposium in the Department of Germanic Languages and Literatures at Washington University, St. Louis, MO.
 Zuberek, S. (2013) The Russian and the Birch Tree - An Exploration of Identity Based on Olga Grjasnowa’s “Der Russe ist einer der Birken Liebt”, The Annual Convention of the Midwest Modern Language Association, Milwaukee, WI.
 Zuberek, S. (2012) Eine Neue Sachlichkeit, die Alte Weiblichkeit – A Critical Interpretation of Gertrud Kolmar’s Novel “A Jewish Mother from Berlin”, The Annual Convention of the South-Central Modern Language Association, San Antonio, TX.
- PROFESSIONAL MEMBERSHIPS:** Linguistic Society of America
 Modern Language Association
 American Association of Teachers of German
 Illinois TESOL – BE