2018

Know thyself? Self- vs. other-assessment of second language pronunciatio

BOSTON UNIVERSITY

SCHOOL OF EDUCATION

Dissertation

KNOW THYSELF?

SELF- VS. OTHER-ASSESSMENT OF

SECOND LANGUAGE PRONUNCIATION

by

MUSHI LI

B.A., Tianjin University of Finance and Economics, 2005
M.A., Tianjin University of Finance and Economics, 2008
M.A., The Ohio State University, 2009

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Education

2018

Approved by

First Reader _____

Marnie Reed, Ed.D.
Clinical Professor of Education

Second Reader _____

Catherine O'Connor, Ph.D.
Dean *ad interim* of the School of Education
Professor of Education and Linguistics

Third Reader _____

Allen G. Harbaugh, Ph.D.
Clinical Assistant Professor of Quantitative Methods

ACKNOWLEDGMENTS

I would like to thank my dissertation committee – Dr. Marnie Reed, Dr. Catherine O'Connor, and Dr. Gregg Harbaugh. They are the most wonderful committee anyone can ask for. Throughout these years, they have helped me tremendously to grow as a researcher and scholar, and I feel so fortunate to have studied under their tutelage.

I would also like to thank my family for being there for me and for raising me to become the person I am today.

Last but not least, I would like to thank everyone who has participated in and helped me with my dissertation project. This would not have been possible without you!

KNOW THYSELF?

SELF- VS. OTHER-ASSESSMENT OF

SECOND LANGUAGE PRONUNCIATION

MUSHI LI

Boston University School of Education, 2018

Major Professor: Marnie Reed, Ed.D., Clinical Professor of Education

ABSTRACT

This dissertation investigates how L2 speakers' assessment of their own pronunciation compares to the assessment of these speakers' pronunciation by different types of listeners.

Study 1 investigated the associations between L2 speakers' pronunciation self-assessment and the assessment by L1 listeners. Eighty-two L2 English speakers performed a picture narrative task and rated their own speech. These speech samples were also rated by eight inexperienced L1 English listeners. Pearson correlation and paired t-test analyses revealed that the speakers' self-assessment was significantly different from L1 English listeners' assessment, and that poor performers overestimated their performance while top performers underestimated it.

Study 2 investigated the associations between L2 speakers' pronunciation self-assessment and the assessment by L1 listeners, L2 listeners who shared an L1 with the speakers, and L2 listeners who did not share an L1 with the speakers. Forty-one L1 Mandarin speakers performed a picture narrative task in English and rated their own pronunciation. These speech samples were also rated by L1 English listeners, L1

Mandarin listeners, and L1 mixed listeners. Pearson correlation and paired t-test analyses revealed that the alignment between self- and other-assessment varied according to the L1 background of the listeners and the construct under evaluation.

Study 3 investigated if L2 listeners had an advantage over L1 listeners at comprehending L2 speech, and if the L1 background and proficiency level of the L2 speakers and listeners affected this potential advantage. Forty-one Mandarin-accented English speech samples from a picture narrative task were rated for comprehensibility by three groups of listeners – L1 English listeners, L1 Mandarin listeners, and L1 mixed listeners. Paired t-test analyses revealed that L1 Mandarin listeners perceived the Mandarin-accented speech to be more comprehensible than the L1 English listeners did, and this benefit was observed with three different proficiency combinations when proficiency was taken into consideration. Although overall the L1 mixed listeners did not perceive the Mandarin-accented speech to be more comprehensible than the L1 English listeners did, when proficiency was taken into consideration, the picture was more complex – while a comprehensibility benefit was observed with one specific proficiency combination, a comprehensibility detriment was observed with a different proficiency pairing.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

GENERAL INTRODUCTION

Introduction

The importance of English proficiency in today's world has changed drastically due to globalization and English becoming a Lingua Franca. It is now vital for speakers of English as a second language (L2) to use English effectively in order to succeed in achieving their goals, whether in their professional or personal lives. Pronunciation, an aspect of language ability, is an essential component of communicative competence (Morley, 1991). For example, English is universally used for international aeronautical radiotelephony communications. Over the past several decades, the NASA aviation system has reported thousands of cases of communication breakdowns between pilots and air traffic controllers, many of which resulted from pronunciation difficulties (NASA, 2014). While not all pronunciation difficulties result in life-threatening situations, the ramifications of issues in pronunciation are real and very serious. Apart from the personal frustration L2 users may feel, numerous studies have demonstrated that pronunciation difficulties can lead to educational, professional, and social consequences (Davila, Bohara, & Saenz, 1993; Lambert, Hodgson, Gardner, & Fillenbaum, 1960; Lev-Ari & Keysar, 2010; Ryan, & Carranza, 1975). Yet despite its apparent importance, it is evidently difficult to change: many L2 learners still suffer from pronunciation difficulties.

Fraser (2010) has noted that for many learners, pronunciation is "simultaneously the most difficult of the language skills and the one they most aspire to master" (p. 358).

Surprisingly, within second language pedagogy, pronunciation, compared to other aspects of second language acquisition research, has been historically under-represented and overlooked. In the past few decades, an increasing amount of attention has been directed to the marginalization of pronunciation within the field of applied linguistics and its subsequent lack of presence in L2 classrooms (Breitkreutz, Derwing, & Rossiter, 2001; Derwing & Munro, 2005; Pennington, & Richards, 1986; Foote, Trofimovich, Collins, & Soler-Urzúa, 2013), which has led Thomson and Derwing (2014) to declare that "the tide has shifted" in L2 pronunciation (p.1).

Nonetheless, despite the extensive ongoing studies exploring the teaching and learning of pronunciation, research still has yet to provide a comprehensive account in regard to L2 speech assessment. One area where more evidence is called for is *how accurately L2 English speakers are able to assess their own English pronunciation*.

For L2 speakers of English, the ability to accurately assess one's own pronunciation skills is of critical importance. It not only affects one's success in English learning, but also one's achievement in professional and personal life. Due to globalization and English becoming a Lingua Franca, today English may be used in various settings (e.g. English as a second language, English as a foreign language, or English as an international language) with different types of interlocutors (e.g. first language English users, L2 English users). Research has shown that pronunciation demand may vary markedly by the communicative setting and when conversing with different interlocutors (e.g. Jenkins, 2002; Bent & Bradlow, 2003). Therefore, learners in the same ESL classroom could have drastically different learning objectives depending

on their respective target communicative contexts. Given this, an increasingly large amount of responsibility is placed on the learners themselves in order to achieve their individualized pronunciation objectives. Without an accurate judgement of their own pronunciation ability, L2 learners may not be able to take charge of their own learning effectively and engage in educational experiences that best facilitate their own learning needs. Additionally, for L2 speech development to take place, great importance has been attached to L2 learners' ability to notice the similarities and differences between their own linguistic output and target form (Gass & Mackey, 2006; Long, 1996; Schmidt, 2001; Schmidt & Frota, 1986). Based on such a hypothesis, having an accurate judgement of one's pronunciation ability may be crucial in L2 pronunciation acquisition considering its role in facilitating objective comparison between one's speech production and external standards. In addition to its essential role in facilitating the learning of L2 pronunciation, L2 speakers' ability to assess their own pronunciation also impacts their professional and personal life. While an inflated self-view may lead to unexpected communication breakdown and cost L2 speakers important opportunities, an overly modest self-view may prevent L2 speakers from pursuing skill-appropriate opportunities and fully taking advantage of the talents they own.

While some studies have explored the associations between L2 English speakers' pronunciation self-assessment and the assessment of these speakers' pronunciation assigned by listeners who speak English as a first language (L1), far less is known about how L2 English speakers' self-assessment would compare to the judgement of listeners who are L2 users of English. There is also a need for more research investigating how

factors such as the L1 backgrounds and proficiency level of L2 interlocutors may affect the associations between self- and other-assessment of L2 pronunciation. Without more evidence-based answers to these questions, language instructors and learners are often left without guidance when making pedagogical and educational decisions.

Study 1 of this dissertation focuses on interactions between L1 and L2 users of English, and investigates whether there is any discrepancy between L2 English speakers' pronunciation self-assessment and L1 English listeners' assessment of these L2 speakers' pronunciation. Study 2 delves into interactions between L2 users of English, and examines how L2 English speakers' pronunciation self-assessment compares to the judgement of listeners who are also L2 users of English. Study 3 intends to offer a close-up examination of the patterns observed in the first two studies by investigating specifically how L2 speech assessment is mediated by speaker and listener L1 backgrounds and proficiency level.

The first problem, how L2 English speakers' pronunciation self-assessment compares to the assessment by L1 English listeners, is of interest because the ability to accurately assess one's own pronunciation skill is essential for L2 learners to achieve their individualized pronunciation objectives. Despite its importance, research targeting L2 pronunciation self-assessment is scarce. In Study 1, a follow-up of a recent study by Trofimovich, Isaacs, Kennedy, Saito, and Crowther (2016[1]) on pronunciation self- vs. other- assessment, a group of L2 speakers of English from a variety of L1 backgrounds rated their own speech from a picture narrative task. Their self-ratings were then

[1] First published online 29 December 2014.

compared to the ratings assigned by L1 English listeners. The primary objective of this study was to test if the results of Trofimovich et al. (2016) held with a different group of subjects and when certain research conditions were modified.

Still focusing on L2 pronunciation self- vs. other- assessment, Study 2 expanded the scope of "other" to include not only L1 English listeners, but also listeners who are L2 users of English. The rationale behind this expansion of scope is the increasing recognition that many L2 speakers are using a shared L2 to communicate with other L2 speakers rather than communicating with L1 speakers, which is particularly true for English (Nelson, 2011). While pronunciation research focusing on L2-L2 interaction is increasing, many of the findings related to pronunciation assessment still assume an L1 English listener (e.g., Foote, 2010, Trofimovich et al., 2016). Study 2 intends to address this complexity of interaction by comparing L1 Mandarin speakers' pronunciation self-ratings to those assigned by L1 English listeners, L1 Mandarin listeners, and listeners whose L1 is neither English nor Mandarin. The primary objective of Study 2 was to investigate the alignment between pronunciation self- and other-assessment when different types of listeners were taken into consideration.

One factor that can potentially affect the alignment between self- and other-assessment of L2 pronunciation is the L1 background of L2 interlocutors. It has long been suggested that L2 speakers may be more intelligible to L2 listeners than L1 listeners (Weinreich, 1953), especially when the L2 speakers and L2 listeners share the same L1. There has been some research investigating whether L2 listeners experience less difficulty understanding L2 speech, and if their ability to comprehend L2 speech is

affected by speaker and listener L2 proficiency. Nonetheless, empirical evidence is very limited. A better understanding of how the language background and proficiency of L2 speakers and listeners impact speech assessment is needed in order to know how to set teaching priorities that will prepare learners to successfully communicate with a wide range of interlocutors. Study 3 is intended to extend prior research in this particular area, and provide explanations and clarifications for the results uncovered in Study 1 and 2. In Study 3, high-proficiency and low-proficiency Mandarin-accented English speech was presented to and rated by L1 English listeners, high-proficiency and low-proficiency L1 Mandarin listeners, as well as high-proficiency and low-proficiency listeners whose L1 is neither Mandarin nor English. Afterward, statistical analyses were carried out to test if L2 listeners indeed rated L2 speech differently from L1 English listeners, and how their particular L1 backgrounds and English proficiency affected their speech assessment.

Overarching Review of the Literature

The Nativeness Principle vs. the Intelligibility Principle

Historically, pronunciation research and pedagogy have been influenced by two contradictory principles, the nativeness principle and the intelligibility principle. Comparing the two, Levis (2005) notes that the nativeness principle holds that "it is both possible and desirable to achieve native-like pronunciation in a foreign language", while the intelligibility principle holds that "learners simply need to be understandable" (p. 370), recognizing that the speaker and listener are both essential elements for communication, and that having an accent, even a strong one, does not necessarily

impede understanding. Since the 1960s, due to the large amount of evidence showing that achieving a native-like accent is an unrealistic goal for those learning a language after puberty (Flege & Frieda, 1995; Moyer, 1999; Piske, MacKay, & Flege, 2001; Scovel, 1969, 1988), the intelligibility principle has gradually replaced the nativeness principle, and became the tenet of pronunciation research and pedagogy.

However, although intelligibility is believed by most L2 researchers and practitioners today to be the goal of pronunciation teaching, the nativeness principle continues to have its presence in the language curriculum and L2 research. In a review of seventy-five L2 pronunciation studies, Thomson and Derwing (2014) found that 63% of the studies implicitly aligned with the nativeness principle. According to the authors, some L2 pronunciation studies targeted pronunciation features that do not likely interfere with intelligibility, such as subphonemic differences or consonants that carry a low functional load, while some other studies used acoustic analyses, which "cannot serve as a replacement for listener judgments" (p. 337), and cannot measure the differences and changes in listener perception. More studies situated within the intelligibility principle are called for in order to provide L2 educators and learners with the necessary information when making pedagogical and educational choices related to pronunciation.

Accentedness, Comprehensibility, and Intelligibility

It is important to contrast three perceptual constructs of pronunciation: accentedness, intelligibility, and comprehensibility. Accentedness describes the extent to which an individual's L2 speech differs from a particular variety of English (Derwing &

Munro, 2005); comprehensibility refers to the amount of perceived listener effort it takes to understand a message (Derwing & Munro, 2009); and intelligibility is defined as the degree of a listener's actual comprehension of an utterance (Derwing & Munro, 2009). As summarized by Derwing & Munro (2009), "accent is about difference, comprehensibility is about the listener's effort, and intelligibility is the end result" (p. 480).

Studies comparing these three constructs have found accentedness, comprehensibility, and intelligibility to be related but partially independent dimensions (e.g. Derwing & Munro, 2009; Derwing, Rossiter, Munro, & Thomson, 2004; Munro & Derwing, 1995; Munro & Derwing, 1999; Trofimovich & Isaacs, 2012; Varonis & Gass, 1982).

In terms of how accentedness may be related or distinct from the other two constructs, generally speaking, those who are low in comprehensibility and intelligibility also tend to be highly accented. However, a strong accent does not necessarily reduce the comprehensibility or intelligibility of L2 speech (Munro & Derwing, 1995), and it is possible for comprehensibility to improve even when there is no noticeable improvement in degree of accentedness (Derwing, Munro, & Wiebe, 1998). Derwing and Munro (2015) offered a discussion of the possible combinations of the extremes of intelligibility and accentedness and the effect these combinations have on the listeners. Using their descriptors, when intelligibility and accentedness are both high, "utterance is fully understood; accent is very strong". When intelligibility is high but accentedness is low, "utterance is fully understood", and "accented is barely noticeable". In the case when

intelligibility is low but accentedness is high, “utterance is not (fully) understood; accent is very strong” (p. 6). And when intelligibility and accentedness are both low, the issue is not relevant to pronunciation anymore, but instead could be related to grammatical issues, words choice, or non-linguistic factors such as noise. Though not stated in Derwing and Munro (2015), a similar set of combination can also be drawn for comprehensibility and accentedness.

The relationship between intelligibility and comprehensibility is closer but these still do not correlate perfectly (Munro & Derwing, 1995, 1999). As noted by Foote (2015), “being able to understand what a person says does not take into account difficulties that a listener may have in processing speech” (p. 6). Derwing and Munro (2015) clarified the distinction between intelligibility and comprehensibility by discussing how different combinations of the extremes of the two constructs may affect the listener. Using their descriptors, if an utterance has high intelligibility and high comprehensibility, “utterance is fully understood; little effort required”. If the intelligibility is high but the comprehensibility is low, “utterance is fully understood”, but “great effort is required”. In the case when intelligibility and comprehensibility are both low, “utterance is not (fully) understood; great effort is exerted”. As for the fourth combination, low intelligibility and high comprehensibility, the authors stated that though probably rare, in this instance the “utterance is not fully understood; however, the listener has the false impression of having easily determined the speaker’s intended meaning” (p. 6).

The means of measuring these three constructs has also been approached

differently in L2 pronunciation research. It has been argued that the evaluation of comprehensibility and accentedness should primarily rely on listener perception given that "what listeners perceive is ultimately what matters most" (Derwing & Munro, 2009, p. 478). Additionally, listener rating has also been suggested to be a reliable approach to assess accentedness and comprehensibility, with very often high intra-class correlations, whether for L1 or L2 listeners (Derwing & Munro, 2009). Conventionally, comprehensibility and accentedness are both measured using Likert-style rating scales.

Intelligibility, on the other hand, is somewhat difficult to assess. In previous studies, various methods have been adopted to assess intelligibility. One of the most common methods is speech transcription, in which listeners hear utterances and write them out in standard orthography (e.g. Bent & Bradlow, 2003; Burda et al., 2003; Derwing & Munro, 1997), and speaker intelligibility is subsequently calculated using the number of words correctly transcribed. Though regarded as the conventional measure of intelligibility, speech transcription has been critiqued as capturing only the individual words spoken, rather than the overall speaker message (Foote, 2015). Other approaches used in previous studies to measure intelligibility include comprehension questions (Anderson-Hsieh & Koehler, 1988), cloze tests (Smith & Rafiqzad, 1979), picture selection in response to a stimulus (Smith & Bisazza, 1982), elicitation of summaries (Perlmutter, 1989), and true/false determination (Munro & Derwing, 1995). Each of these approaches has strengths and drawbacks, and no one method is fully adequate as a measure for intelligibility (Munro, Derwing, & Morton, 2006).

The measure of comprehensibility was used in all three studies of this

dissertation. Comprehensibility was selected over intelligibility to measure the broad understanding of L2 pronunciation based on the following three factors. One factor is the practical consideration that comprehensibility, which is generally measured by rating scales, reflects a more practical and realistic approach compared to intelligibility, which is typically assessed via speech transcription (Trofimovich & Isaacs, 2012). Secondly, intelligibility is often operationalized narrowly, "equating understanding with simply being able to identify the actual words spoken rather than the message being conveyed" (Foote, 2015, p. 5). Thus comparatively, as far as communication is concerned, comprehensibility captures the concept of understanding in a more relevant way. An L2 speaker who "is able to speak words that can be identified, but whose speech is difficult to process and whose meaning does not come across easily, is likely to struggle with interlocutors despite having a high level of intelligibility" (p. 8). Additionally, intelligibility, in its broad sense, refers to listeners' general ability to understand speech, which is not usually distinguished from comprehensibility (Levis, 2006). As discussed in Isaacs and Trofimovich (2012), this broad sense of intelligibility is commonly adopted in L2 testing, such as the Test of English as a Foreign Language (TOEFL) or the International English Language Testing System (IELTS), in which intelligibility, rated via rating scales, was in fact comprehensibility. Therefore, the construct of comprehensibility in the current dissertation falls under Levis' (2006) broad sense of intelligibility and thus "reflects a typical approach to assessing intelligibility in oral proficiency scales" (Isaacs & Trofimovich, 2012, p. 477).

In addition to comprehensibility, the accentedness measure was also included in

Study 1 and Study 2 of this dissertation. The inclusion of accentedness intends to serve two purposes. Firstly, the goal of pronunciation teaching and learning is generally believed to be comfortable intelligibility and comprehensibility. Given that accentedness, comprehensibility, and intelligibility are independent constructs, and that a strong accent may not impede understanding on the listeners' part (Derwing, Rossiter, Munro, & Thomson, 2004; Munro & Derwing, 1995), the inclusion of accentedness can "enhance our knowledge of the nature of foreign accents and their effects on communication" (Derwing & Munro, 2005, p. 379). Such knowledge is informative for educators and learners in setting teaching and learning priorities. Additionally, since acquiring a native-like accent is very rare for L2 learners who start learning an L2 after early childhood (Flege, Munro, & Mackay, 1995; Scovel, 2000), it has been suggested that educators would be doing learners a disservice by not reminding them to be realistic with their desire to attain a native-like accent (Derwing & Munro, 2005). At the same time, research has also informed us that a perfectly native accent may be attainable for a very small number of highly motivated (Moyer, 2004) individuals with special aptitude (Ioup, Boustagi, El Tigi, & Moselle, 1994), and it has been argued that learners should not be denied this possibility if that is what they truly want (Harmer, 2001). Therefore, it is believed that a better understanding on self-assessment of accentedness could potentially be informative to these particular learners. This being said, the current studies follow the intelligibility principle, and the results on accentedness will be discussed within a framework where comfortable intelligibility and comprehensibility are the primary concerns.

Theoretical Frameworks

All three studies of this dissertation were conducted within the intelligibility principle as opposed to the nativeness principle. Additionally, Study 1 and 2 were motivated by the Interaction Hypothesis (Long, 1996) and the Noticing Hypothesis (Schmidt, 2001; Schmidt & Frota, 1986). For second language development to take place, it is important for learners to notice the differences between their own speech and that of their interlocutors. The awareness of a "gap" between one's interlanguage and the target form is believed to be essential for L2 acquisition (Gass & Mackey, 2006; Schmidt, 2001; Schmidt & Frota, 1986). The term interlanguage is used to refer to the linguistic system that contains features from both the learners' native language and the language that is being learned (Selinker, 1972). The importance of noticing the similarities and differences between one's own linguistic performance and the interlocutor's language is also highlighted in interaction-driven learning. According to the Interaction Hypothesis (Long, 1996), the process of interacting with another individual leads to negotiation of meaning, which serves to draw the learners' attention to these gaps between their interlanguage and the target form. The first two studies of this dissertation do not focus on the interaction between speakers per se, but rather embrace the principle that for a pronunciation feature to be potentially acquired, it has to be noticed by the learners somehow.

CHAPTER TWO

SELF- VS. OTHER-ASSESSMENT OF SECOND LANGUAGE PRONUNCIATION

Introduction

The ability to accurately assess one's skills is crucial to the success of individual lives and the society as a whole. People's self-perception guides decision-making – while accurate self-perception increases the chance of effective and appropriate decision-making, flawed self-assessment is more likely to lead people to make potentially skill-inappropriate choices, or miss out on opportunities to fully utilize the talents they truly own. The importance of English proficiency in today's world has changed drastically due to globalization and English becoming a Lingua Franca. For learners of English as an L2, the ability to accurately assess one's own pronunciation skills is of critical importance. An accurate assessment of one's pronunciation ability enables English learners to engage in educational and communicative experiences that are appropriate for their skill levels, which not only enhances communicative success, but also promotes acquisition of the target language.

First, accurate self-assessment promotes learner-centered learning and learner autonomy (Little, 2005; Oscarson, 1989; Rogerson-Revell & Miller, 1994). Today, comfortable intelligibility (whether learner speech can be understood) and comprehensibility (the amount of effort it takes to understand a message in the perception of a listener) have replaced "native-like" speech, and have become the goal of

pronunciation teaching and learning (Celce-Murcia et al., 2010; Levis, 2005). English may be used in settings such as ESL (English as a Second Language – a context where English is the primary language of the country), EFL (English as a Foreign Language – a context where English is not an official language of the country and residents do not regularly use English to communicate with each other), or EIL (English as an International Language), also known as ELF (English as a Lingua Franca – a context where English functions as the medium of communication between speakers who do not share the same L1). The goals of present-day pronunciation teaching take both the speaker and listener into account, and therefore pronunciation demands vary by the setting, which may result in highly varied and individualized learning objectives (Jenkins, 2002). L2 English learners studying in the same classroom may be learning English with the goal of being able to communicate with different interlocutors in different settings after the completion of the course - while it may be necessary for some to acquire English pronunciation that is intelligible to L1 English speakers in an ESL setting, some may only need to be comprehensible to fellow L2 English speakers from the same L1 background. Considering such variations, an increasingly large amount of responsibility will be placed on the learners themselves in setting goals and directing their own learning. Achieving individualized pronunciation objectives relies on accurate self-assessment.

Secondly, the ability to accurately assess one's pronunciation skills may be critical for the acquisition of pronunciation features. According to the Interactionist Theory, the process of interacting with another individual leads to negotiation of meaning, which serves to draw the learners' attention to a "gap" between his/her

interlanguage and the target form. The awareness of such a gap is believed to be essential for L2 acquisition (Gass & Mackey, 2006; Schmidt & Frota, 1986). Therefore, if faulty self-assessment presents itself in L2 pronunciation, learners may not be able to perceive the similarities or differences between their own L2 production and the target form, and may thus fail to benefit from such approaches as anticipated.

Self- vs. Other-assessment

Despite the importance of accurate self-perception, people's self-assessment and their actual performance are often poorly correlated (Carter & Dunning, 2008; Dunning, Heath, & Suls, 2004; Zell & Krizan, 2014). As Benjamin Franklin once stated, "there are three things extremely hard: steel, a diamond, and to know one's self". Numerous studies in various skill domains have suggested that people's perception of their own competence often diverges markedly from their actual competence. For example, judges' confidence in detecting deception correlated very weakly with their actual accuracy (DePaulo, Charlton, Cooper, Lindsay, & Muhlenbruck, 1997). What consumers thought they knew about their purchases did not align well with what they actually knew (Alba & Hutchinson, 2000). Resident physicians' self-ratings of their interpersonal skills when communicating with patients only agreed at a low level with their patients' opinions (Millis et al., 2002). The meta-analysis Mabe and West (1982) conducted revealed the correlation between self-perception of knowledge and objective performance is only .29. Within higher education, Falchikov and Boud (1989) found that on average students' self-assessment was only moderately related to the assessment assigned by their teachers.

It has been suggested that the accuracy of one's self-perception varies across skill

domains. There appears to be an increase in self-assessment accuracy when the skill domain is specific and clearly defined, and when the performance tasks are objective, familiar, or low in complexity (Burson, Larrick, & Klayman, 2006; Dunning, Meyerowitz, & Holzberg, 1989; Hayes & Dunning, 1997; Zell & Krizan, 2014). For example, people tend to have a more accurate judgment of their own note-taking skills compared to general academic ability (Dunning et al., 1989). Football athletes are better at assessing their sports ability than their ability to detect lies (Zell & Krizan, 2014).

The Dunning-Kruger Effect

What has been called the "Dunning-Kruger" effect has been observed in many social and intellectual domains; it captures the tendency for poor performers to over-evaluate themselves, and for top performers to under-estimate themselves (Kruger & Dunning, 1999). The tendency among poor performers to over-estimate themselves has particularly been noted extensively in a wide range of skill domains. The inability for poor performers to accurately assess themselves has been proposed to be due to a double curse: "Not only do these people reach erroneous conclusions and make unfortunate choices, but their incompetence robs them of the metacognitive ability to realize it" (Kruger & Dunning, 1999, p.1121). Poor performers were not the only ones who reach erroneous self-appraisals. Top performers have been found to underestimate their ability and test performance relative to their peers due to their assumption that their proficiency is shared by their peers (Burson, Larrick, & Klayman, 2006; Kruger & Dunning, 1999).

L2 Pronunciation Self-assessment

Given what we know about people's self-assessment ability, what should be expected in terms of people's ability to assess their own L2 pronunciation skills?

First, pronunciation, compared to other aspects of second language abilities, is a complex skill. It encompasses multiple dimensions, from individual sounds to the overall intonation, and it is "the only aspect of language that has a neuromuscular basis", requires "neuromotor involvement", and has a "physical reality" (Scovel, 1988, p.101). Secondly, pronunciation remains an ill-defined skill. In the past decades, the goal of pronunciation teaching has experienced paradigm shifts, which resulted in a nonuniformity of the concept of pronunciation. While an increasing population is viewing pronunciation competence as comfortable intelligibility and comprehensibility, it is still believed by many that the goal of pronunciation acquisition is the accent of a native speaker (Tokumoto & Shibata, 2011). Thirdly, L2 speakers typically lack the necessary information which enables accurate assessment of their pronunciation skills. Despite the growing recognition in the past decades, pronunciation is still rarely a focus of instruction in L2 classrooms, which resulted from both its incompatibility with the communicative teaching approach, and a lack of teacher training in this specific area (Breitkreutz, Derwing, & Rossiter, 2009; Celce-Murcia, Brinton, & Goodwin, 2012; Foote, Holtby, & Derwing, 2011; Foote, Trofimovich, Collins, & Soler-Urzúa, 2013; Pennington & Richards, 1986). The absence of pronunciation instruction leads to learners' lack of diagnostic abilities and metalinguistic knowledge, particularly at lower ability levels (Derwing, 2003; Derwing & Rossiter, 2002). Moreover, it has been well established that

recasts (repeating the erroneous production back to the learner in a corrected form), the most commonly adopted form of feedback, are low in salience and may not be recognized by learners as correction (Lyster, 2001).

Finally, pronunciation difficulties are highly context-dependent and individualized. Listeners differ in their tolerance for foreign accent (Moyer, 2013), and their ability to comprehend accented speech (Grant, 2014). Additionally, different pronunciation issues may be associated with communication difficulties for native speaker (NS) – non-native speaker (NNS) conversation vs. NNS-NNS communication (Jenkins, 2002). In sum, considering its complexity, ill-defined nature, context-dependency, and the lack of necessary information in the learning and communicative environment, it may be predicted that English learners are susceptible to having an inaccurate assessment of their L2 English pronunciation.

Studies have examined the associations between self- and other-assessment of some aspects of L2 ability, primarily receptive skills such as listening and reading. While a few studies reported positive associations between self-assessment and external assessment (Brantmeier & Vanderplank, 2008; Krausert, 1991; LeBlanc & Painchaud, 1985), the majority of studies reported poor associations (e.g. Brantmeier, 2006; Davidson & Henning, 1985; Edele, Seuring, Kristen, & Stanat, 2015; Falchikov & Boud, 1989; Janssen van Dieten, 1989).

In comparison, research targeting L2 pronunciation self-assessment is scarce. At present, there is limited evidence regarding L2 speakers' ability to evaluate their own pronunciation skills. Several studies have examined the accuracy of L2 learners' self-

assessment of their overall L2 pronunciation skill (Derwing, 2003; Foote, 2010; Lappin-Fortin & Rye, 2014; Raasch, 1980), and a few teased apart the linguistic domains underlying pronunciation, and investigated learners' ability to assess their own production or reception of segmental sounds or prosody (Dlaska & Krekeler, 2008; Lappin-Fortin & Rye, 2014; Yule, Damico, & Hoffman, 1987).

Yule, Damico, and Hoffman (1987) examined L2 users' ability to assess their own pronunciation receptive skills. Focusing on intermediate-level English learners' perception of segmental sounds (e.g., cloud vs. crowd), the authors examined the accuracy of these learners' perception in a listening task and their self-monitoring ability measured by a confidence-rating scale at two times separated by seven weeks. Across all three groups of students (lower, middle, and higher proficiency according to their initial test accuracy scores), little correlation was observed between the learners' self-confidence rating of accuracy and their actual accuracy at both times of observation, with the lower group improving only in perception accuracy, and the middle group improving only in self-monitoring ability.

A few studies focused on self-assessment of pronunciation productive skills.

Lappin-Fortin and Rye (2014) investigated the potential value of self-assessment in an intermediate university French pronunciation course. Using the reading of the same passage in the beginning (pre-test) and at the end of a course (post-test), students were asked to make a global assessment of their pronunciation in the pre-test (using the question - "what score out of ten would you give yourself for this reading?"), and judge specific aspects of their pronunciation (which were studied in the course) in the post-test.

Comparing students' self-ratings to the ratings by experienced L1 French raters, the results indicated that students were relatively accurate in their judgment in both the pre-test and post-test, although they showed the tendency to overestimate the extent to which their abilities were native-like in certain segmental aspects and prosody. This study demonstrated the potential value of integrating self-assessment in L2 pronunciation courses in helping learners acquire L2 pronunciation, but it did not evaluate L2 speech in terms of the actual intelligibility and comprehensibility.

Dlaska and Krekeler (2008) investigated L2 learners' ability to detect differences between their own segmental production and the target form. Advanced learners of German were asked to compare the recording of their own German segmental production to a native-speaker model, and indicate if they found their own production to be the same or different as the native speaker model. The results revealed that the learners were only able to recognize 44% of the inaccurate sounds identified by experienced native raters. While the study highlighted L2 learners' difficulty detecting differences between their own production and the target form, it implicitly aligned with the nativeness principle.

In another study which focused on the accentedness measure, Foote (2010) recorded L2 speakers' readings of two sentences ("Young children can be very noisy", and "Many people drink coffee for breakfast"), and compared the accentedness ranking assigned by L2 speakers themselves and by L1 English raters. The results revealed that L2 speakers' self-assessment of accentedness not only did not have a positive correlation with that of L1 English raters on a significant level, but occasionally showed a slightly negative correlation with L1 English raters' ratings.

The only study that included measures that are in line with the intelligibility principle was Trofimovich et al. (2016), who investigated if there existed a discrepancy between L2 speakers' self-assessment and experienced native listeners' ratings in terms of comprehensibility and accentedness. As the results show, the L2 English speakers in their study had mostly inaccurate self-assessment of how accented and comprehensible their speech sounded. The study also investigated the Dunning-Kruger effect specifically, and found evidence that was consistent with the Dunning-Kruger effect, such that the poor performers in their study over-estimated themselves while the top performers underestimated themselves.

A review of the literature exploring self- vs. other-assessment of L2 pronunciation shows that existing studies are not only small in number, but also vary drastically in their target linguistic features and study designs. A majority of these studies were not carried out under the intelligibility principle, thus are unable to provide evidence in terms of how understandable L2 speech was actually perceived to be. Additionally, many of these studies used speech from text reading for speech assessment. Text reading (also referred to as read aloud) can be an advantageous method of speech elicitation in many circumstances, especially when more control over content, syntactic structure, and vocabulary use is necessary. However, to examine how understandable L2 speech is in real-life settings, extemporaneous speech may provide us with more relevant information. So far, there is only one study (Trofimovich et al., 2016) in L2 pronunciation self-assessment that was conducted within the intelligibility principle and that utilized extemporaneous speech. Additionally, in Trofimovich et al. (2016), the judgement of

experienced L1 English raters was used as the base of comparison when evaluating self-assessment accuracy. The experienced raters in their study all had L2 teaching experience, held advanced degrees in applied linguistics or language teaching, and had completed at least one course in applied phonetics and pronunciation teaching. While the utilization of experienced L1 English raters captured how L2 speech may be assessed in a typical ESL classroom, the results may not be generalizable to real-life settings where L2 English speakers often communicate with L1 English speakers who do not have an ESL background, especially considering existing research evidence which shows that experienced and inexperienced L1 raters may judge L2 speech differently (e.g. Kennedy & Trofimovich, 2008; Saito & Shintani, 2016; Thompson, 1991).

In sum, more evidence is needed to gain a fuller understanding of L2 speakers' ability to assess their own L2 pronunciation.

Experienced vs. Inexperienced L1 Raters

In L2 pronunciation studies, the term "experienced raters", also named "expert raters", or "experienced judges", has been used to refer to phoneticians and speech therapists (Cucchiarini, Strick, & Boves, 2002), English as a second language (ESL) teaching assistants (Calloway, 1980), English as a foreign language teachers (Bongaerts, van Summeren, Planken, & Schils, 1997), and experienced ESL teachers who either held or were pursing postgraduate degrees in applied linguistics (Isaacs & Thomson, 2013). In contrast, those who do not meet these criteria have been referred to as non-expert, inexperienced raters, naïve raters, novice rater, or "person in the street" (Thompson, 1991, p. 177).

One area of interest in L2 pronunciation research is whether inexperienced L1 listeners rate L2 speech differently from experienced L1 raters. Some studies reported no difference between the two groups. For example, Bongaerts et al. (1997) found no significant difference between experienced and inexperienced (based on the criterion of presence or absence of linguistic training) raters' accent ratings. Similarly, Calloway (1980) found the two groups agreed substantially on their ratings of accentedness and comprehensibility. Both Derwing et al. (2004) and Isaacs and Thomson (2013) reported no group difference on the ratings of fluency, accentedness, and comprehensibility between these two types of L1 English raters. On the other hand, other studies reported that raters who have more relevant background with accented speech tend to judge L2 speech more leniently (e.g. Kennedy & Trofimovich, 2008; Saito & Shintani, 2016; Thompson, 1991). For example, Thompson (1991) found that experienced raters were more lenient in their ratings of accentedness compared to inexperienced raters. Winke, Gass, and Myford (2013) investigated if raters' L2 learning experience may be a potential source of bias in their ratings of TOEFL iBT test takers' oral performance. The findings suggested that raters with Spanish L2 learning experience were significantly more lenient with L1 Spanish test takers, as were L2 Chinese raters with L1 Chinese test takers. Saito and Shintani (2016) compared Canadian and Singaporean raters' ratings of comprehensibility using L2 speech samples collected from picture description tasks, and found that the Singaporean raters, "who not only used various models of English but also spoke a few L2s on a daily basis in a multilingual environment", assigned significantly higher comprehensibility scores to Japanese-accentedness speech samples. The authors

attributed the findings to the Singaporean raters' "relatively high sensitivity to, in particular, lexicogrammatical information" (p. 421). Similarly, Kennedy and Trofimovich (2008) investigated how previous exposure to non-native speech impacted the perception of intelligibility, comprehensibility, and accentedness of L2 speech, and found that ESL teachers understood more speech from both L1 and L2 speakers than the listener group that reported having had little to no contact with L2 speakers of English, probably due to the ESL teachers' "greater knowledge of how L2 speakers' pronunciation differs from that of native speakers" (p. 478).

In sum, though inconclusive, research evidence has shown that raters' L2 background, including accent familiarity, linguistic training experience, and L2 learning experience, significantly affects their assessment of L2 speech, arguably because the exposure to highly variable stimuli promotes perceptual learning and adaptation to foreign-accented English (Bradlow & Bent, 2008). Therefore, if raters have more L2 experience, they are likely to "judge L2 audio tokens more leniently and analyze their own rating processes more clearly than inexperienced and novice raters" (Saito & Shintani, 2016, p. 423).

The Current Study

The current study is a follow-up to Trofimovich et al. (2016), and reexamined the associations between L2 speakers' self-assessment of their pronunciation in relation to L1 English listeners' assessment of these L2 speakers' pronunciation. The primary goal of this study is to test if the findings of Trofimovich et al. (2016) held with a different group

of L2 English speakers when inexperienced L1 English raters were utilized instead of experienced L1 raters. It is believed that the selection of inexperienced English L1 raters reflects a setting that may be more realistic and more likely to represent the environment in which English is used for these L2 English speakers.

Another difference between the current study and Trofimovich et al. (2016) concerns the rating procedure. In order to minimize the variability in self-assessment due to methodological differences between how speakers and listeners assess speech, the current study adjusted the rating procedures in Trofimovich et al. (2016), and made the L2 speakers' self-assessment procedure more similar to that of the listeners. In the current study, the L2 speakers listened to their own narratives prior to self-rating. Additionally, both the speakers and listeners in the current study listened to and rated the same three speech samples (accentedness level roughly ranged from high to low) as a practice before proceeding to speech rating.

The specific research question asked is: Is there any discrepancy between L2 English speakers' self-assessment and L1 English listeners' assessment of the accentedness and comprehensibility of these L2 speakers' English speech?

Method

Participants

Speaker

The speakers in the study were eighty-two L2 English users (57 female, 25 male) with a mean age of 21.5 years (SD = 4.1) from 18 L1 backgrounds, including Mandarin

Chinese (41), Arabic (16), Japanese (6), Spanish (5), German (2), French (2), Thai (2), Creole (1), Korean (1), Kinyarwanda (1), Polish (1), Swahili (1), Kalenjin (1), Italian (1), Kazakh (1), Portuguese (1), Greek (1), and Vietnamese (1). One subject speaks both Arabic and French as L1s, one Swahili and Kalenjin as L1s, and one German and Polish as L1s. All three subjects were counted twice in the L1 counts. All speakers were enrolled in a university (54) or English language school (28) in the northeast U.S. during the time the study was carried out. Out of the subjects enrolled in a university, twelve were in graduate programs and forty-two in undergraduate programs. They arrived in the U.S. to pursue studies at a mean age of 20.9 (SD=4), and had been studying in the U.S. for a mean of 8.9 months (SD=12.4).

Sixty-nine speakers had recently taken either TOEFL iBT (57 subjects) or IELTS tests (12 subjects), which are high-stakes instruments used to assess the participants' English ability to pursue studies in English-speaking higher education institutes. The participants' mean scores were 97.7 (SD = 13.7) for TOEFL iBT overall, 25.3 (SD=4.1) for TOEFL iBT listening, and 23.4 (SD=3.4) for TOEFL iBT speaking. The TOEFL listening and speaking sub scores were missing from three participants. The participants' mean scores were 6.17 (SD=0.94) for IELTS overall, 6.46 (SD=1.28) for IELTS listening, and 6.29 (SD=.78) for IELTS speaking. Among all reported TOEFL scores, five subjects had scores that were over 2 years old. Among the IELTS scores reported, two subjects had IELTS scores that were over 2 years old.

Excluding one participant who left out the second page of the language background questionnaire, eighty-one participants had studied English for an average of

11.6 years (SD = 4), primarily through formal instruction in primary, secondary, and university-level settings. These eighty-one speakers also self-rated their English ability at a mean of 5.88 (SD = 1.38) in speaking and 6.33 (SD = 1.65) in listening using 9-point Likert-type scales (1 = extremely poor, 9 = extremely good). Using a 0–100% scale (0% = never, 100% = all the time), they also estimated their daily use of English at 61.36% (SD=23.66%).

Listeners

This group included eight inexperienced L1 English listeners (4 male, 4 female), with a mean age of 20.4 (SD = 3.7). All raters were L1 speakers of North American English, who reported using English an average of 98.6% (SD = 1.7%) of time in their daily life, out of which approximately 96.9% (SD = 3.6%) of time was spent interacting with other L1 speakers of English (as opposed to L2 English speakers). None of the subjects had any ESL/EFL experience, or prior linguistic training experience. Given that all these raters were residing in Boston during the time of the study, a city with a large international population, the exposure to and familiarity with foreign accents was considered acceptable. Accents that the subjects reported familiarity with to varying levels include German, French, Spanish, Italian, and Japanese.

Procedure

Speakers

During the individual research meeting held with each speaker, the participant first filled out a questionnaire, which collected information such as age, gender, years of English learning, TOEFL iBT or IELTS score, years of residence in an English-medium

country, age arriving in an English-speaking country, etc. The participants were also instructed to submit a copy of their TOEFL or IELTS score report if available. In several cases, a score report was not available, and self-reported scores were accepted.

Afterward, each participant performed a picture narrative task. An eight-frame picture was used for speech elicitation, which depicts two travelers bumping into each other and accidentally exchanging their identical suitcases (see figure 1 below). Having first appeared in a study by Derwing, Munro & Thomson (2008), this picture has been used for speech elicitation in a number of L2 pronunciation studies, and was selected here for the purpose of cross-study consistency and comparison.

Figure 1. Image used for speech elicitation.

Each subject was presented with the picture sequence and instructed to narrate a story describing what happened in each image. There was no time limit imposed for preparation and narration. The narratives were recorded directly onto a computer and

stored as digital audio files. Upon completing the narrative task, the subject used a 9-point scale to indicate how well they performed the task (1 = very poorly, 9 = very well) and to estimate overall task difficulty (1 = very easy, 9 = very difficult).

Next, each participant received a training session. The training session was to prepare the speakers to rate the comprehensibility and accentedness of their speech in the picture narrative task. The constructs of comprehensibility and accentedness were explained to the participants, and three practice speech files that are not relevant to the story depicted in the picture narrative task were subsequently played for rating practice. Each participant was invited to ask any question that they might have in regard to the two constructs. Afterward, the speakers used a 9-point scale to indicate how well they understood the two concepts (1 = don't understand at all, 9 = understand very well), and how comfortable they were at using these constructs to rate their own speech (1 = not comfortable at all, 9 = completely comfortable). In the end of the meeting, they were played the recording of their own narrative and self-rated the accentedness and comprehensibility of their own speech.

After the meetings with all eighty-two L2 English speakers were completed, the narrative recordings were turned into stimulus items before being presented to listeners for speech rating. All recordings were normalized for volume by matching peak amplitude across files. They were then edited to remove all fillers and false starts at the beginning of the file and shortened to include only the initial 30 seconds of speech, consistent with prior research using 20-60 seconds samples to evaluate speech (e.g., Derwing et al., 2004).

Listeners

The meetings with the L1 English listeners took place after all meetings with L2 English speakers were completed. During the individual meeting with each L1 English listener, he/she first completed a language background questionnaire, then received the same training session and proceeded to speech rating. Given the much larger sample of Mandarin-accented speech, the speech samples collected from the 41 Mandarin speakers were divided in two groups, 21 speech files in group 1 and 20 in group 2. The speech samples from the rest of the speakers were also divided into two groups, 21 in group 3 and 20 in group 4. Each group was rated by four randomly chosen L1 English listeners for accentedness and comprehensibility. The rationale behind the division of speech samples was to facilitate an even distribution of speech samples with different accents for each L1 English listener.

The participants were instructed that they could play and rate the recordings at their own pace, with an unlimited number of replays permitted, which is consistent with Trofimovich et al. (2016). Each participant was also told that although they were not required to listen to the entire recording to make a decision, they had to listen to at least 15 seconds of each recording, which is consistent with 15-30 seconds' samples used to obtain listeners' impressionistic ratings of speech in prior research (e.g. Derwing et al., 2004). Additionally, the participants were also informed that all recordings were cut off after 30 seconds, and once they completed the ratings for a recording, they may not go back and change their assigned ratings.

For the rating of accentedness and comprehensibility, 9-point Likert-type numerical scales were used as the instrument for speech measurement (see Figure 2). The rationale behind the scale length lies in that a 9-point scale is capable of capturing the magnitude of accent that raters may detect (Southwood & Flege, 1999), that it is safer to "overestimate the listeners' ability to resolve accentedness than to underestimate it" (Munro & Derwing, 1994, p. 259), and that it "make results comparable to other studies" (Isaacs & Thomson, 2013, p. 137). For accentedness, 1 denotes heavily accented, and 9 not accented at all. For comprehensibility, 1 indicates hard to understand, and 9 easy to understand.

Accentedness: This refers to how much a speaker's speech is influenced by his/her native language and/or is colored by other non-native features.

1	2	3	4	5	6	7	8	9
heavily accented								not accented at all

Comprehensibility: This refers to how much effort it takes to understand what someone is saying. If you can understand with ease, then a speaker is highly comprehensible. However, if you struggle and must listen very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility.

1	2	3	4	5	6	7	8	9
hard to understand								easy to understand

(adapted from Saito & Shintani, 2016)

Figure 2. 9-point scales used for speech rating.

Analysis and Results

Cronbach's alpha was computed across the eight L1 English listeners' ratings in order to measure inter-rater reliability, separately for accentedness and comprehensibility. The obtained coefficients were 0.95 for accentedness and 0.89 for comprehensibility, both of which exceed the benchmark value of .70 - .80 (Larson-Hall, 2010). Fleiss' κ was run to determine if there was agreement between the L1 English listeners' judgements in terms of how accented and comprehensible each speech sample was. At an agreement window of ±1, κ (accent) = 48.2% and κ (comprehensibility) = 38.0%. At an agreement window of ±2, κ (accent) = 69.1% and κ (comprehensibility) = 61.5%. The Fleiss' κ values, taken together with the Cronbach's alphas, provide strong evidence that there is coarse inter-rater reliability. Therefore, a single accentedness and comprehensibility score was derived for each speaker by averaging across the 4 ratings assigned by L1 English listeners.

Consistent with Trofimovich et al. (2016), for each speaker, an overconfidence score was also calculated by subtracting the mean L1 English listener rating from the speaker's self-rating, separately for comprehensibility and accentedness. A positive overconfidence score represented speakers' overestimation of their accentedness or comprehensibility relative to the judgement of L1 English listeners, while a negative score indicated an underestimate of their own accentedness or comprehensibility. A score around zero indicated self-ratings that were aligned with listener assessments.

The first set of analyses examined the relationship between the L2 speakers' actual performance (as rated by L1 English listeners) and their self-ratings. Pearson

correlation tests (one-tailed) revealed moderate associations between the speakers' self-ratings and L1 English listener ratings for accentedness, $r(80) = .46$, $p < .0001$, and moderate associations for comprehensibility, $r(80) = .57$, $p < .0001$. Results from paired samples T-test analyses indicated significant differences between speakers' self-ratings of accentedness ($M = 5.79$, $SD = 1.83$) and their performance rated by L1 English listeners ($M = 5.2$, $SD = 2.07$), $p = .011$, Cohen's d (effect size) = .3. Significant differences were also detected between self-ratings of comprehensibility ($M = 6.93$, $SD = 1.64$) and the ratings assigned by L1 English listeners ($M = 6.54$, $SD = 1.59$), $p = .019$, Cohen's d (effect size) = .25. Overall, although speaker self-ratings and L1 English listener ratings were moderately associated, they were found to be significantly different from each other, with speakers over-evaluating themselves in both accentedness and comprehensibility.

Given that evidence consistent with the Dunning-Kruger effect was reported in Trofimovich et al. (2016), the next set of analyses set out to investigate if the Dunning-Kruger effect can be found again in the current study. Results from Pearson correlation tests show that there were moderate associations between speakers' overconfidence scores and their actual performance, both for accentedness, $r(80) = –.32$, $p = .0017$ (one-tailed), and comprehensibility, $r(80) = –.44$, $p < .0001$ (one-tailed). The negative associations observed here indicate that more accented and less comprehensible speech (as perceived by L1 English listeners) was associated with greater overconfidence (illustrated in Figure 3). In other words, the speakers who were perceived by L1 English listeners as more accented and less comprehensible were those who over-estimated their

own ability, a pattern that is consistent with the Dunning-Kruger effect and the results in Trofimovich et al. (2016).

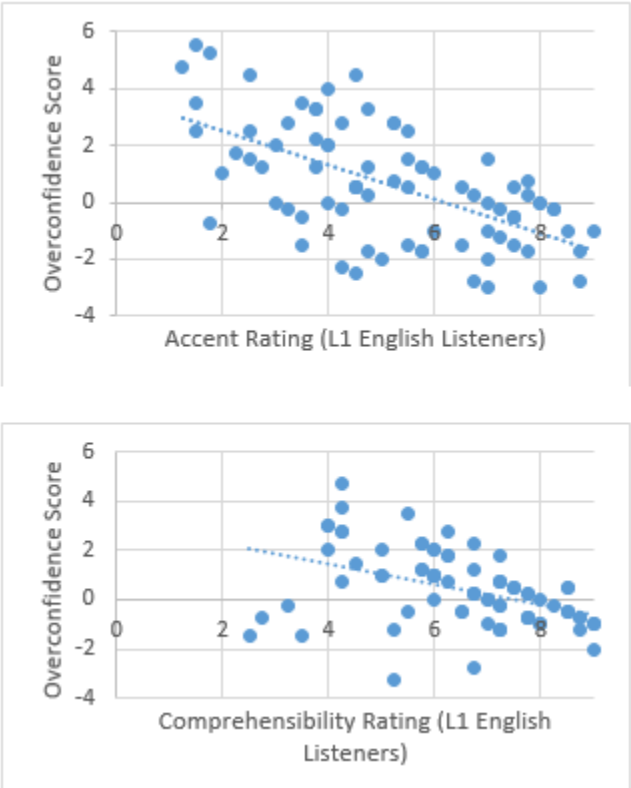

Figure 3. Associations between L2 speakers' (n = 82) overconfidence scores and their actual performance (as rated by L1 English listeners) for accent (top) and comprehensibility (bottom), with regression lines showing the best fit to the data.

In line with Trofimovich et al. (2016), the overconfidence scores of the bottom and top thirds of the speakers were compared. For accentedness, the bottom thirds include 27 subjects, and the accentedness scores (rated by L1 English listeners) covered in this bottom group ranged from 1.25 to 4. The top thirds (28 subjects) covered accentedness scores 6.5-9. For accentedness, the bottom thirds of the speakers were

significantly more overconfident (M = 2.12, SD=1.87) than the top thirds (M = –.81, SD=1.19), who were under-confident, $p < .0001$, Cohen's d (effect size) = 1.92.

For comprehensibility, 24 subjects were categorized in the bottom third group (comprehensibility score 2.5-5.75) and 26 in the top third group (comprehensibility score 7.5-9). Again, the bottom third group was significantly more overconfident (M = 1.24, SD=1.96) than the top third (M = –.48, SD=.67), who underestimated their performance, $p < .001$, d = 1.31.

To present the Dunning-Kruger effect in a more straightforward fashion, the accentedness and comprehensibility ratings (by both L2 speakers themselves and L1 English listeners) were first rank-ordered, then expressed as percentile scores, an approach that is in line with Trofimovich et al. (2016). The relationship between the percentile-based measure of the speakers' actual and self-rated performance is illustrated in Figure 4, where speakers' self-rated percentile rankings (solid orange line) and the L1 English listener ratings (dashed blue line) were plotted separately for the four speaker groups (bottom, second, third, and top quartile) based on L1 English listeners' ratings. As shown in Figure 4, the L2 English speakers who were rated by L1 English listeners as the bottom 25th percentile overestimated their performance (self-ratings higher than L1 English listener ratings), while the L2 speakers in the top 25th percentile underestimated their own performance (self-ratings lower than L1 English listener ratings). Self-ratings and L1 English listener ratings were in fact only aligned for speakers whose performance was around 50th percentile for both accentedness and comprehensibility, which roughly

corresponded to 5.25 for accentedness and 6.88 for comprehensibility (L1 English listeners rating on a scale of 1-9).

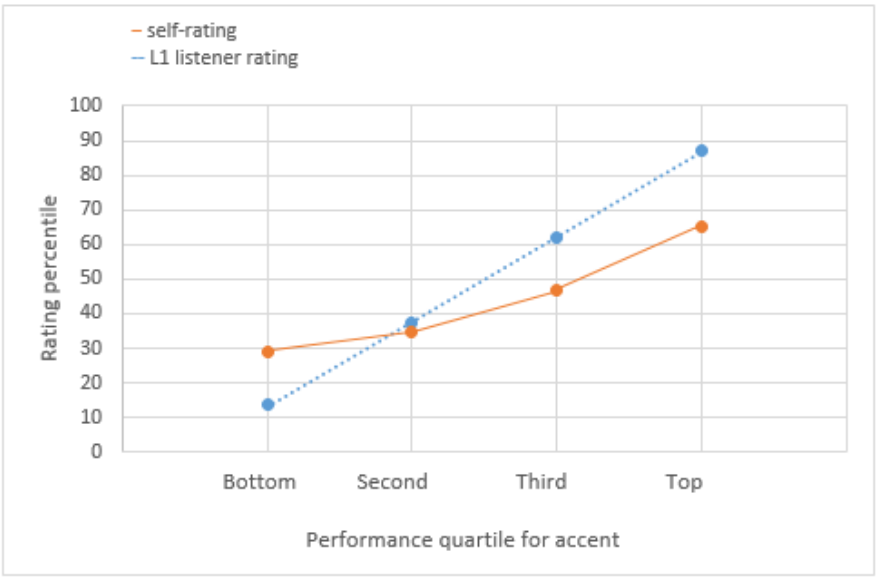

Figure 4. L2 speakers' (n=82) percentile rankings for self- and L1 English listener-ratings of accentedness (top) and comprehensibility (bottom) as a function of L1 English listener-rated performance quartile (bottom to top 25%).

The next set of analyses investigated how various speaker variables correlated with their overconfidence scores. These background characteristics include age, gender, L1, length of time studying in the U.S, length of time spent in English-speaking countries, age of first exposure to English, TOEFL scores, TOEFL listening subscores, TOEFL speaking subscores, self-rated speaking and listening ability, amount of daily English use, as well as the speakers' self-rated task difficulty and task performance success (both rated after the speakers completed the picture narrative task).

Among all these characteristics, the only ones that were found to be significantly correlated with L2 speakers' overconfidence scores were TOEFL speaking and listening scores, age, and L1. Among the subjects who reported TOEFL scores, significant associations were found between TOEFL listening and accentedness overconfidence scores, $r(52) = -.28$, $p=.02$, and between TOEFL listening and comprehensibility overconfidence scores, $r(52) = -.31$, $p=.01$. TOEFL speaking subscores were also found to be significantly associated with speakers' accentedness overconfidence scores, $r(52) = -.52$, $p<.0001$, and between TOEFL speaking scores and comprehensibility overconfidence scores, $r(52) = -.47$, $p<.001$. The negative correlations indicated that a decrease in the performance in TOEFL listening and speaking was associated with an increase in their overconfidence scores in accentedness and comprehensibility. These results are consistent with the Dunning-Kruger effect found in the current study as reported above.

Significant associations were found between age and comprehensibility overconfidence scores, $r(80) = .26$, $p=.018$. This finding is unexpected since it is

generally believed for expertise to advance as a person ages. Therefore, given the Dunning-Kruger effect observed in this study and numerous previous studies, an increase of age is expected to be associated with a decrease in overconfidence scores. The reason behind the findings here is unclear, but given that .26 is a rather small correlation, there is the possibility that this may be a Type I error. More studies investigating age and L2 pronunciation overconfidence are needed.

To test if the speakers' L1 background was associated with their overconfidence scores, speakers were categorized in L1 Mandarin vs. L1 non-Mandarin groups. The reason behind this categorization is that Mandarin represents the largest L1 group within the current sample (n=41). Spearman's correlation was run to assess the relationship between L1 and accentedness overconfidence scores. A weak correlation was detected, which was statistically significant, r_{pb} = -.28, *rs* =.27, p= 0.014. This indicates that L1 Mandarin speakers, in comparison to L1 speakers of non-Mandarin, were associated with a higher degree of overconfidence in terms of how accented their speech sounded.

To further investigate the pattern uncovered here, the largest two L1 groups, L1 Mandarin (n=41) and L1 Arabic (n=16) were taken out to test if there were any significant associations between L1 and accentedness overconfidence scores. The results from Spearman's correlation tests revealed no significant association between L1 and accentedness overconfidence scores for these two L1s, r_{pb} = -0.1, *rs* = 0.062, p=0.6478.

The associations between accentedness overconfidence and L1 (L1 Mandarin vs. L1 neither Mandarin nor Arabic) were subsequently tested, and moderate correlations were detected between L1 and accentedness overconfidence scores, r_{pb} = -0.35635, *rs* =

0.39, p= 0.0013. The significant correlations here indicate that L1 Mandarin speakers, in comparison to those who are neither L1 Mandarin nor L1 Arabic speakers, were associated with a higher degree of overconfidence in the accentedness of their own speech. The reason behind the significant findings here is unclear. Given that on average, L1 Mandarin L1 speakers (M=4.82) were more accented than the speakers who were neither L1 Mandarin nor L1 Arabic speakers (M=5.6) as perceived by L1 English listeners, the differences in their English proficiency may be a contributor to the differences observed here. However, given that no significant association was found between overconfidence scores and L1 for the L1 Mandarin vs. L1 Arabic pair despite the comparable accentedness level of the L1 Arabic (M=5.6) and L1 neither Mandarin nor Arabic group (M=5.6), it may be speculated that there are other factors at play here that could also have had an impact on these speakers' self-assessment ability. Given the small number of subjects representing different L1s in the current study, future studies are needed to better understand how L1 background may impact L2 pronunciation self-assessment.

Other than the factors discussed above, there was no significant association between any of the other characteristics and overconfidence scores, $r<.22$, $p>.05$.

What's worth mentioning here is that among these various background characteristics, gender was not found to be significantly associated with the speakers' overconfidence scores. This is an interesting finding particularly considering existing evidence which suggested that males and females tend to differ in their self-assessment behavior (Dunning et al., 2004; Pallier, 2003). However, given the gender imbalance in

the current study (57 female, 25 male), future research is needed to further investigate the impact of gender on L2 pronunciation self-assessment.

To sum up, although L2 speakers' self-assessment of comprehensibility and accentedness were moderately associated with the assessment by L1 English listeners, the ratings assigned by these two groups were significantly different from each other. Consistent with prior studies (e.g., Trofimovich et al., 2016), L2 speakers' self-ratings, compared with their actual performance (as rated by L1 English listeners), reflected the Dunning Kruger effect, with speakers at the bottom of the accentedness and comprehensibility scale overestimating their performance while speakers at the top of each scale underestimating it.

Discussion

The current study investigated whether there was any discrepancy between L2 speakers' own assessment and L1 English listeners' assessment of these speakers' L2 accentedness and comprehensibility. Consistent with previous research on L2 speakers' pronunciation self-assessment ability (Trofimovich et al., 2016) and the Dunning-Kruger effect (Carter & Dunning, 2008), the L2 speakers in this study showed mostly inaccurate self-assessment of how accented and comprehensible they sounded, relative to the ratings of inexperienced L1 English listeners. Consistent with prior findings (Trofimovich et al., 2016), speakers at the low end of the accentedness and comprehensibility scales overestimated their performance, while speakers at the high end of each scale underestimated it.

A Comparison with Trofimovich et al. (2016)

However, while the study by Trofimovich et al. (2016) revealed tenuous correlation between self- and L1 English listener-assessment (For accent, r=.06, and for comprehensibility, r=.18), the current study revealed moderate associations that were significant for both accentedness and comprehensibility. One speculation of the difference observed here is that it may be related to the type of L1 English listeners used in the two studies. Compared to the experienced L1 English raters used in Trofimovich et al. (2016), the inexperienced listeners utilized in the current study may perceive L2 accent and comprehensibility differently. However, since the current study did not include a group of experienced L1 English listeners as a comparison group, it is unknown if the type of L1 English listeners is indeed the cause of the differences in the strength of correlation.

Another speculation regarding the observed differences in the strength of correlation is that it may have stemmed from methodological differences between Trofimovich et al. (2016) and the current study. In Trofimovich et al. (2016), while the listeners had the opportunity to compare one speaker to another, and engage in norm-referenced assessment, the speakers had no access to their own performance, nor were they given a chance to review a reference sample of other speakers to mediate their use of the rating scales prior to self-assessment. In contrast, the speakers and listeners in the current study listened to and rated the same three L2 speech samples (accentedness level roughly ranged from high to low) during the training session, and the speakers were also played their own speech recording prior to self-assessment. These changes in

methodology made the self-assessment procedure more similar to that of the listeners', and the three practice speech samples could have mediated the use of the rating scales for both L1 Mandarin and L1 English participants, both of which could have contributed to a higher level of agreement between self- and other-assessment. Future studies that specifically compare these different rating procedures are needed to confirm this speculation.

Nonetheless, the moderate correlations found in the current study should be understandable given the years of formal English learning experience the subjects had. With feedback on their English ability regularly being provided, it is to be expected that these L2 English learners have a general sense where their English ability falls within the overall proficiency spectrum.

Inaccurate Self-assessment

For second language development to take place, it is important for learners to notice the differences between their own speech and that of their interlocutors. According to the Interaction Hypothesis (Long, 1996), the process of interacting with another individual leads to negotiation of meaning, which serves to draw the learners' attention to these gaps between their interlanguage and the target form. However, if L2 speakers do not have an accurate impression of their own L2 ability, as revealed by the current study, they may have difficulty comparing their own speech to others', and thus may be unable to notice how their own speech differs from their interlocutors'.

The current study also revealed patterns in line with the Dunning Kruger effect. This finding has implications for L2 learning, such that, as discussed in Trofimovich et

al. (2016), L2 speakers with lower ability might be overconfident in their self-assessment, "making it harder for them to notice their linguistic shortcomings". In comparison, those "at the higher end of the spectrum, who are conservative in their self-assessment, might preoccupy themselves with linguistic issues which are fairly inconsequential to their performance" (p. 134). Taken together, L2 learners' lack of accuracy to assess their own pronunciation ability may result in unwarranted self-confidence or an overly modest self-view, both of which could cost L2 speakers important opportunities.

Calibration of Self-assessment with Objective Ability

Despite the evident lack of accuracy, the ability to accurately assess one's own pronunciation skill is of critical importance for the success of an L2 learner. Accurate pronunciation self-assessment enables L2 learners to engage in educational and communicative experiences that are appropriate for their skill levels, which not only enhances communicative success, but also promotes the acquisition of the target language.

Various factors may affect the accuracy of self-assessment. Greater accuracy has been observed when the skill domain is specific and clearly defined, and when the performance tasks are objective, familiar, or low in complexity (Burson, Larrick, & Klayman, 2006; Dunning, Meyerowitz, & Holzberg, 1989; Hayes & Dunning, 1997; Zell & Krizan, 2014). Research has shown that when the domains under evaluation were ambiguous and open to the use of idiosyncratic criteria and evidence, people provided self-serving appraisals which tend to diverge from objective evaluations (Dunning, Meyerowitz, & Holzberg, 1989; Hayes & Dunning, 1997). Up until today, pronunciation

remains an ill-defined skill for many L2 educators and learners. It would be beneficial for L2 educators to obtain a better understanding of the constructs of pronunciation and the objectives of pronunciation teaching, increase the amount of pronunciation instruction in their classroom, integrate effective pronunciation pedagogy that is supported by empirical evidence, and make an active effort to "demystify" pronunciation to L2 learners. Such effort will help clarify the constructs and criteria of L2 pronunciation, and equip students with necessary metalinguistic knowledge about pronunciation, which may enhance their diagnostic ability and self-assessment accuracy.

Additionally, inaccurate self-assessment has been attributed to a lack of crucial information needed to reach objective opinions (Carter & Dunning, 2008; Dunning, Heath, & Suls, 2004). Feedback tends to be biased, flawed, or missing (Carter & Dunning, 2008). Positive feedback is often withheld, while negative feedback is often disguised. Moreover, given that feedback is often probabilistic in real life, the outcome can be inconsistent with the quality of choices and behaviors (Carter & Dunning, 2008). Given such, educators should provide L2 learners with easily recognizable negative feedback about their skills and abilities. It has been well established that recasts (repeating the erroneous production back to the learner in a corrected form), the most commonly adopted form of corrective feedback in L2 classrooms, are low in salience and may not be recognized by learners as correction (Lyster, 2001). Additionally, as a part of corrective feedback, educators should also provide explicit information in terms of why failure has occurred. As pointed out by Kruger and Dunning (1999), failure is subject to more attributional ambiguity than success. For success to occur, "many things must go

right: The person must be skilled, apply effort, and perhaps be a bit lucky"; whereas for failure to occur, "the lack of any one of these components is sufficient" (p. 1882). Because of this, even if people receive feedback that points to a lack of skill, they may attribute it to some other factors (Snyder, Higgins, & Stucky, 1983; Snyder, Shenkel, & Lowery, 1977), thus fail to learn about the true level of their abilities.

In sum, pronunciation is a complex and ill-defined skill. Speakers "might succeed in communication despite a noticeable accent or through the use of such strategies as gesturing, avoidance, or circumlocution to convey a message, and without interlocutors' feedback focusing specifically on speech perception and production" (Trofimovich et al., 2016, p. 134). With all these factors potentially contributing to a lack of accuracy in L2 speakers' pronunciation self-assessment skill, it is essential for L2 educators to provide the clarification and information necessary to reduce the ambiguity L2 learners may experience with L2 pronunciation.

Conclusion and Future Directions

In conclusion, the current study investigated if there was any discrepancy between L2 English speakers' pronunciation self-assessment and inexperienced L1 English listeners' assessment of these L2 speakers' English pronunciation. Consistent with Trofimovich et al. (2016), the L2 speakers in the current study perceived the accentedness and comprehensibility of their pronunciation to be significantly different from the judgement of L1 English raters. Given the importance of having an accurate assessment of one's own L2 pronunciation, it is essential for L2 educators to include the

calibration of L2 pronunciation self-assessment as a teaching objective, which may be facilitated by clarifying the constructs and expectations of pronunciation to L2 learners and providing the learners with complete and unambiguous feedback on their performance. The findings of the current study extended the literature on self-assessment in social, academic, and professional domains.

Both the original study by Trofimovich et al. (2016) and the current study utilized picture narrative tasks for speech elicitation. Prior research has suggested that task type and task difficulty may be associated with the level of accuracy in self-assessment (Bachman & Palmer, 1989; Heilenmann, 1990). It is suggested that future studies investigating L2 pronunciation self-assessment use additional task types with varying degrees of cognitive demand in order to allow a better understanding of whether the results of the current study may be generalized when other types of tasks are utilized. Additionally, in the current study there was an imbalance within the subjects in terms of gender, age, and cultural background. It may be important for future studies to examine how these social-psychological factors may be linked to overconfidence, given that males and females may differ in their self-assessment behaviors (Dunning et al., 2004; Pallier, 2003), and that socially-construed norms may encourage or discourage overconfidence (Fay, Jordan, & Ehrlinger, 2012; Matsuno, 2009).

In terms of how L2 learners may be assisted to calibrate their self-assessment, a few different methods were suggested in Dunning et al. (2004), which include review of past performance, benchmarking (comparing self-performance against that of others), and peer assessment. Future studies are encouraged to examine how these different methods

may effectively help calibrate L2 learners' pronunciation self-assessment.

Additionally, the findings from the current study suggest that L2 English speakers from different L1 backgrounds may assess their own English pronunciation differently. Future studies that specifically examine the impact of L1 background on L2 pronunciation self-assessment are needed to gain a better understanding in this respect.

Last but not least, it is important for future studies to examine the possible contributions of various linguistic factors in L2 speakers' inaccurate self-assessment. Trofimovich et al. (2016) found that when making comprehensibility assessment, L2 speakers and L1 English listeners took different linguistic variables into consideration - while L2 speakers' judgement of their own comprehensibility was linked only to segmental, suprasegmental, and fluency variables, L1 English listeners' ratings were linked to additional factors such as lexicon, grammar, and discourse structure. Such findings indicated that L2 speakers may be unaware which linguistic factors make L2 speech comprehensible to listeners. A better understanding in this direction will provide L2 educators useful information in terms of how to adjust instructional foci in order to help their students better understand and improve L2 comprehensibility.

CHAPTER THREE

SELF- VS. OTHER-ASSESSMENT OF SECOND LANGUAGE PRONUNCIATION – A COMPARISON ACROSS DIFFERENT LISTENER TYPES

Introduction

For learners of English as a second language (L2), the ability to accurately assess one's own pronunciation skills is of critical importance. Due to globalization and English becoming a Lingua Franca, today English may be used in various settings, such as English as a second language, English as a foreign language, or English as an international language, or with different types of interlocutors, such as first language (L1) English users or L2 English users. Research has shown that pronunciation demand may vary markedly by the communicative setting and when conversing with different types of interlocutors. Therefore, an increasingly large amount of responsibility is placed on the learners themselves in order to achieve their individualized pronunciation objectives. Without an accurate judgement of their own pronunciation ability, L2 learners may not be able to take charge of their own learning effectively and engage in skill-appropriate educational experiences that best facilitate their own learning needs. Additionally, for L2 speech development to take place, great importance has been attached to L2 learners' ability to notice the similarities and differences between their own linguistic output and the target form (Gass & Mackey, 2006; Long, 1996; Schmidt, 2001; Schmidt & Frota, 1986). This means having an accurate judgement of one's own pronunciation ability may

be crucial in L2 pronunciation acquisition considering its role in facilitating objective comparison between one's speech production and external standards. Moreover, L2 users' accurate self-assessment of their pronunciation also enhances their professional and personal life. For example, in the scenario of an important job interview, L2 users' inflated self-view in their pronunciation skill may lead to inadequate preparation and communication breakdown without their awareness, which may cost them good opportunities. On the other hand, an overly modest self-view may prevent L2 users from pursuing such opportunities entirely, thus fail to take full advantage of the talents they truly own.

So then, how well do L2 speakers assess their own L2 pronunciation? A number of studies have investigated the accuracy of L2 learners' pronunciation self-assessment (Foote, 2010; Lappin-Fortin & Rye, 2014; Trofimovich et al., 2016). However, these studies compared L2 speakers' self-assessment to the judgement of L1 listeners rather than L2 listeners. For example, Foote (2010) compared the accentedness ranking determined by speakers themselves to the ranking assigned by L1 English raters. Lappin-Fortin and Rye (2014) compared French L2 learners' self-ratings to the ratings assigned by experienced L1 French raters. In another study by Trofimovich, Isaacs, Kennedy, Saito, and Crowther (2016), L2 English speakers' self-assessment of the accentedness and comprehensibility of their own English speech was compared to the judgement of experienced L1 English listeners.

While L2 research commonly makes the assumption that the goal of learning the L2 is to communicate primarily with L1 speakers of the language, the reality in today's

world is that the total number of L2 users of English has surpassed that of L1 users (Crystal, 2003). This means many are learning English to communicate with interlocutors who are also L2 users of English. It has been argued that for those who are using English in an English as an International Language (EIL) setting, making their own speech understandable to L2 English listeners may be the more important goal than adapting to native speaker norms (Jenkins, 2002). However, given our limited understanding in how L2 speech is judged by L2 listeners, "additional work comparing the responses of native speakers (NSs) and non-native speakers (NNSs) is needed to develop a more complete understanding of L2 speech intelligibility" (Derwing & Munro, 2005, p. 382). Therefore, as far as L2 speakers' pronunciation self-assessment ability is concerned, it is not only important to understand how L2 speakers assess their own speech in relation to L1 English listeners' perspective, but also how their self-assessment compares to the judgement by L2 users of English.

Role of L1 Background in L2 Assessment

In today's world, English may be used in various settings (e.g. ESL, EFL, EIL) with different types of interlocutors (e.g. L1 English users, L2 English users). Kachru (1997) proposed the concentric circles model to capture how English is used in different parts of the world: the Inner Circle, the Outer Circle, and the Expanding Circle. Today, as people constantly move around the globe and interact with each other, different speaker-listener interaction patterns can be observed frequently: NS- NS, NS-NNS, NNS-NS, and NNS-NNS, which may even expand to a nine-square matrix when taking into consideration English speakers from the Inner, Outer, and Expanding Circles (Levis,

2005). As the paradigm in pronunciation research and teaching is switching away from the nativeness principle and towards the intelligibility principle (Levis, 2005), any discussion of pronunciation teaching and learning should take the specific contexts into consideration.

In addition to L1 English listeners' assessment of L2 English pronunciation, how L2 English listeners judge L2 English speech is also of importance due to the increasing recognition that many L2 speakers are using English to communicate with other L2 English speakers rather than L1 speakers (Jenkins, 2002). While some studies have shown that L1 and L2 listeners' judgment of L2 accentedness, intelligibility, and comprehensibility can be quite comparable (e.g. Flege, 1988; Munro, Derwing, & Morton, 2006), plenty of evidence has suggested otherwise - that L2 listeners may perceive L2 speech differently from L1 listeners (e.g. Bent & Bradlow, 2003; Hayes-Harb et al., 2008; Imai et al., 2005; Smith, Bradlow, & Bent, 2003; Winters & O'Brien, 2013). While L1 listeners may find L1 speech more intelligible than L2 speech, the opposite may be true for L2 listeners - ''A [non-native] speaker who cannot make himself understood when speaking English to a native English speaker will have no difficulty conversing in English with another [non-native] speaker'' (Nash, 1969, p. 4).

A Shared L1 Benefit

There is a general belief that L2 users who share an L1 have an advantage understanding each other when communicating in an L2. Research in various fields has investigated the possibility of a shared-L1 intelligibility advantage. From the perspective of cross-language speech perception, such a shared-L1 intelligibility advantage is based

on the principle that "L2 accents are primarily characterized by transfer from the L1", therefore when a listener shares an L1 with the speaker, he or she will have "an intimate familiarity with the phonological patterns of that speaker's L2 accent" (Harding, 2011, p. 165).

However, studies investigating a shared-L1 intelligibility benefit have reported inconclusive findings. In a study by Major et al. (2002), while the L1 Spanish listeners showed a small intelligibility advantage for L1 Spanish speakers, the L1 Chinese listeners actually showed an intelligibility disadvantage when listening to L1 Chinese speakers. Hayes-Harb, Smith, Bent, and Bradlow (2008) investigated the intelligibility of Mandarin-accented English for L1 English and L1 Mandarin listeners. Using a word identification task (minimal pairs that demonstrate word-final voicing contrast, such as "cub" and "cup"), the authors reported that the L1 Mandarin listeners were on average more accurate than the L1 English listeners at identifying words produced by L1 Mandarin speakers. Imai et al. (2003) compared the ability of L1 English listeners and L1 Spanish listeners at recognizing English words produced by an L1 Spanish speaker. The results revealed that the L1 Spanish listeners outperformed L1 English listeners in the word recognition task. Similarly, in an attempt to investigate if L2 speakers' L1s affect their judgment of L2 speech, Munro, Derwing, and Morton (2006) found that Japanese-accented English was more intelligible to L1 Japanese listeners compared to L1 English listeners.

The Current Study

The current study further examined the associations between L2 pronunciation self- and other-assessment by expanding the scope of "other" to include L2 users of the language as well. Given that L2 speakers who share an L1 may have an advantage understanding each other, and that L2 English speakers from different L1 backgrounds may approach pronunciation self-assessment differently, the current study selected L2 English speakers from L1 Mandarin background, and compared their self-assessment to the assessment assigned by L1 English listeners, L2 English listeners who are also L1 speakers of Mandarin, and L2 English listeners who do not speak Mandarin as their L1. By examining the relationships between L1 Mandarin speakers' assessment and the assessment by listeners from a variety of L1 backgrounds, the current study intends to offer a comprehensive examination of L2 English speakers' ability to assess their own L2 pronunciation in various communicative contexts.

The specific research questions are:

1) Regarding the comprehensibility and accentedness of L2 English speech, what are the associations between the self-assessment of L1 Mandarin speakers and the assessment assigned by listeners who are L1 speakers of English?
2) Regarding the comprehensibility and accentedness of L2 English speech, what are the associations between the self-assessment of L1 Mandarin speakers and the assessment assigned by listeners who are also L1 speakers of Mandarin?
3) Regarding the comprehensibility and accentedness of L2 English speech, what are the associations between the self-assessment of L1 Mandarin speakers and

the assessment assigned by listeners who are neither L1 English speakers nor L1 Mandarin speakers?

Method

Participants

Speakers

The speakers were forty-one L1 Mandarin speakers (34 female, 7 male), with a mean age of 20.7 years (SD=3.7), who were enrolled in a university (39) or English language school (2) in the northeast U.S. during the time the study was carried out. Among those enrolled in a university, eleven were in graduate programs, twenty-eight in undergraduate programs. They had arrived in the U.S. to pursue studies at a mean age of 20 (SD=3.8), and have been studying in the U.S. for a mean of 8.6 months (SD=13.4). All speakers had recently taken either TOEFL iBT (39 subjects) or IELTS tests (2 subjects), which are high-stakes instruments that were used to assess the participants' ability to pursue university studies. The participants' mean overall scores were 100.3 (SD = 10.2) for TOEFL iBT, 25.6 (SD=4.0) for TOEFL listening, and 23.3 (SD=2.8) for TOEFL speaking. The TOEFL listening and speaking sub scores were missing from one participant. The participants' mean overall scores were 6.75 (SD=1.1) for IELTS, 7.8 (SD=1.8) for IETLS listening, and 5.8 (SD=.4) for IETLS speaking. Among the TOEFL scores collected, four subjects had scores that were over 2 years old; the IELTS scores from both subjects who reported their IELTS scores were over 2 years old.

Excluding one participant, who accidentally left out the second page of the language background questionnaire, forty participants had studied English for an average of 12.9 years (SD = 2.8), primarily through formal instruction in primary, secondary, and university-level settings. The speakers self-rated their English ability at a mean of 5.6 (SD = 1.4) in speaking and 6.3 (SD = 1.6) in listening using 9-point Likert-type scales (1 = extremely poor, 9 = extremely fluent). Using 0–100% scales (0% = never, 100% = all the time), they also estimated their daily use of English at 57.5% (SD=22.7%).

Listeners

L1 English listeners

This group included eight inexperienced L1 English listeners (4 male, 4 female), with a mean age of 20.4 (SD = 3.7). All raters were L1 speakers of North American English, who reported using English an average of 98.6% (SD = 1.7%) of time in their daily life, out of which approximately 96.9% (SD=3.6%) of time was spent interacting with other L1 speakers of English (as opposed to L2 speakers). None of these subjects had studied Mandarin, and reported low familiarity with Mandarin-accented English (a self-rating that is equal to or smaller than 3 on a 1-9 scale, 1 – not at all familiar, 9 – very familiar). None of the subjects had any ESL/EFL experience, or prior linguistic training experience. Given that all these raters were residing in Boston during the time of the study, a city with a large international population, the exposure to and familiarity with accents other than Mandarin was considered acceptable. Accents that the subjects reported familiarity with to varying levels include German, French, Spanish, Italian, and Japanese.

L1 Mandarin listeners

This group included thirty-eight out of the forty-one participants in the speaker group. This is because two research meetings were scheduled for each participant in the speaker group (L1 Mandarin group). During the first meeting the L1 Mandarin participants served as the speakers, and during the second meeting they served as the listeners. Thirty-eight out of the initial 41 participants returned for the second research meeting.

L1 Mixed listeners

The L1 mixed group were forty-one speakers (23 female,18 male) from an L1 background that is neither Mandarin nor English, with a mean age of 22.3 years (SD=4.3). They were enrolled in a university (15) or English language school (26) in the northeast U.S. during the time the study was carried out. Among those enrolled in a university, one was in a graduate program, fourteen in an undergraduate program. This group includes L1 speakers of Arabic (16), Creole (1), Korean (1), Thai (2), Kinyarwanda (1), Spanish (5), German (2), French (2), Polish (1), Swahili (1), Kalenjin (1), Italian (1), Kazakh (1), Portuguese (1), Greek (1), Japanese (6), and Vietnamese (1). One subject speaks both Arabic and French as L1s, one Swahili and Kalenjin as L1s, and one German and Polish. All three subjects were counted twice in the L1 counts. The participants had studied English for an average of 10.4 years (SD = 4.7), primarily through formal instruction in primary, secondary, and university-level settings. They arrived in the U.S. to pursue studies at a mean age of 21.7 (SD=4.2), and had been studying in the U.S. for a mean of 9.1 months (SD =11.4). Twenty-eight out of the forty-

one subjects had taken either TOEFL iBT (18 subjects) or IELTS tests (10 subjects). The participants' mean overall scores were 92.1 (SD = 18.2) for TOEFL iBT, 24.8 (SD=4.5) for TOEFL listening, and 23.7 (SD=4.5) for TOEFL speaking. The TOEFL listening and speaking sub scores were missing from two participants. The participants' mean overall scores were 6.1 (SD=1.0) for IELTS, 6.2 (SD=1.3) for IETLS listening, and 6.4 (SD=.8) for IETLS speaking. Among the TOEFL scores, one subject had scores that were over 2 years old; none of the IELTS scores was over 2 years old. The speakers self-rated their English ability at a mean of 6.2 (SD = 1.3) in speaking and 6.4 (SD = 1.7) in listening using 9-point Likert-type scales (1 = extremely poor, 9 = extremely fluent). Using 0–100% scales (0% = never, 100% = all the time), they also estimated their daily use of English at 65.1% (SD=24.3%).

Procedure

Speakers

During the first meeting with the L1 Mandarin group, each participant first filled out a questionnaire, which collected information such as age, gender, years of English learning, TOEFL iBT or IELTS score, years of residence in an English-medium country, age arriving in an English-speaking country, and etc. The participants were also instructed to submit a copy of their TOEFL or IETLS score report if available. In several cases, a score report was not available, and self-reported scores were accepted. Afterward, each participant performed a picture narrative task. An eight-frame picture was used, which depicts two travelers bumping into each other and accidentally exchanging their identical suitcases (see Figure 5 below). Having first appeared in a study

by Derwing, Munro & Thomson (2008), this picture was used in a number of L2 pronunciation studies, and was selected here for the purpose of cross-study consistency and comparison.

Figure 5. Image used for speech elicitation.

The subjects were presented with the picture sequence and instructed to narrate a story of what happened in each image. There was no time limit imposed for preparation and narration. The narratives were recorded directly onto a computer and stored as digital audio files. Upon completing the narrative task, the subjects used a 9-point scale to indicate how well they performed the task (1 = very poorly, 9 = very well) and to estimate the overall task difficulty (1 = very easy, 9 = very difficult).

Next, each participant received a training session. The training session was to prepare the speakers to rate the comprehensibility and accentedness of their speech in the picture narrative task. The constructs of comprehensibility and accentedness were

explained to the participants, and three practice speech files that are not relevant to the story depicted in the picture narrative task were subsequently played for rating practice. Participants were invited to ask any question that they might have in regard to the two constructs. Afterward, the speakers used a 9-point scale to indicate how well they understood the two concepts (1 = don't understand at all, 9 = understand very well), and how comfortable they were at using these constructs to rate their own speech (1 = not comfortable at all, 9 = completely comfortable). At the end of the first meeting, they were played the recording of their own narratives and self-rated the accentedness and comprehensibility of their own speech.

After the meetings with all forty-one L1 Mandarin speakers were completed, the narrative recordings were turned into stimulus items before being presented to listeners for speech rating. All recordings were normalized for volume by matching peak amplitude across files. They were then edited to remove all fillers and false starts at the beginning of the file and shortened to include only the initial 30 seconds of speech, consistent with prior research using 20-60 seconds samples to evaluate speech (e.g., Derwing et al., 2004).

Listeners

L1 English listeners

First, each participant completed a language background questionnaire. Afterward, they received the same training session and proceeded to speech rating. The speech samples collected from the forty-one Mandarin speakers were divided into two groups, 21 speech files in group 1 and 20 in group 2. Each group was rated by four

randomly chosen L1 English listeners for accentedness and comprehensibility. The participants were instructed that they could play and rate the recordings at their own pace, with an unlimited number of replays permitted, which is consistent with Trofimovich et al. (2016). Each participant was also informed that although they were not required to listen to the entire recording to make a decision, they had to listen to at least 15 seconds of each recording, which is consistent with 15-30 seconds' samples used to obtain listeners' impressionistic ratings of speech in prior research (e.g. Derwing et al., 2004). Additionally, the participants were also informed that all recordings were cut off after 30 seconds, and once they completed the ratings for a recording, they may not go back and change their assigned ratings.

L1 Mandarin listeners

Thirty-eight out of the initial forty-one L1 Mandarin subjects returned several months later and participated in the second research meeting, where they served as listeners. During the second meeting, first they received the same training session as a review. Then, each participant listened to the 40 speech samples produced by the other Mandarin speakers (41 minus the one produced by the subject him/herself). These listeners received the same speech rating instructions as the L1 English listeners.

L1 mixed listeners

First, each participant completed a language background questionnaire. Afterward, the participants received the same training session, followed by speech rating. They rated the speech samples produced by all forty-one L1 Mandarin speakers. The requirements of the speech rating procedure were the same as the other two listener

groups discussed earlier.

For the rating of accentedness and comprehensibility, 9-point Likert-type numerical scales (Figure 6) were chosen as the instrument for speech measurement. The rationale behind the scale length lies in that a 9-point scale is capable of capturing the magnitude of accent that raters may detect (Southwood & Flege, 1999), that it is safer to "overestimate the listeners' ability to resolve accentedness than to underestimate it" (Munro & Derwing, 1994, p. 259), and that it "make results comparable to other studies" (Isaacs & Thomson, 2013, p. 137). For accentedness, 1 denotes heavily accented, and 9 not accented at all. For comprehensibility, 1 indicates hard to understand, and 9 easy to understand.

Accentedness: This refers to how much a speaker's speech is influenced by his/her native language and/or is colored by other non-native features.

1	2	3	4	5	6	7	8	9
heavily accented								not accented at all

Comprehensibility: This refers to how much effort it takes to understand what someone is saying. If you can understand with ease, then a speaker is highly comprehensible. However, if you struggle and must listen very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility.

1	2	3	4	5	6	7	8	9
hard to understand								easy to understand

(adapted from Saito & Shintani, 2016)

Figure 6. 9-point scales used for speech rating.

Analysis and Results

Cronbach's alpha, a measure of inter-rater reliability, was computed across the English listeners' ratings, separately for accent and comprehensibility. The obtained coefficients were 0.95 for accent and 0.89 for comprehensibility, exceeding the benchmark value of .70 - .80 (Larson-Hall, 2010). Fleiss' κ was run to determine if there was agreement within the L1 English listeners' judgement in terms of how accented and comprehensible each speech sample sounds. At an agreement window of ±1, κ (accent) = 48.2% and κ (comprehensibility) = 38.0%. At an agreement window of ±2, κ (accent) = 69.1% and κ (comprehensibility) = 61.5%. The Fleiss' κ values, taken together with the Cronbach's alphas, provide strong evidence that there is coarse inter-rater reliability. Therefore, a single accentedness and comprehensibility score was derived for each speaker by averaging across the ratings assigned by the four randomly-chosen L1 English listeners.

Cronbach's alpha was also computed across the ratings assigned for each Mandarin-accented speech sample by all L1 Mandarin listeners and by all L1 mixed listeners, separately for accentedness and comprehensibility. For the L1 Mandarin listeners' ratings, the obtained coefficients were 0.99 for accentedness and 0.97 for comprehensibility. For the L1 mixed listeners' ratings, the obtained coefficients were 0.98 for accentedness and 0.97 for comprehensibility. Since the alphas all exceeded the benchmark value of .70 - .80 (Larson-Hall, 2010), a single accent and comprehensibility score was derived for each speaker by averaging across all ratings assigned by all L1 Mandarin listeners, and a single accent and comprehensibility score was derived for each

speaker by averaging across all ratings assigned by the L1 mixed listeners.

The first set of analyses examined the relationships between the L1 Mandarin speakers' performance rated by L1 English listeners and the speakers themselves. Pearson correlation tests (one-tailed) revealed moderate associations between the speakers' self-rated scores and L1 English listener-ratings for accentedness, $r(39) = .54$, $p = .00013$, and moderate associations for comprehensibility, $r(39) = .53$, $p = .00018$. Results from paired samples t-test analyses indicated significant differences between speakers' self-ratings of accentedness ($M = 5.96$, $SD = 1.75$) and their performance rated by L1 English listeners ($M = 4.82$, $SD = 2.22$), $p = .00052$, Cohen's d (effect size) = .58, as well as between self-ratings of comprehensibility ($M = 6.91$, $SD = 1.36$) and the comprehensibility ratings assigned by L1 English listeners ($M = 6.45$, $SD = 1.49$), $p = .036$, Cohen's d (effect size) = .33. Therefore, overall, although speaker self-ratings and L1 English listener-ratings were moderately associated, they were found to be significantly different from each other, with speakers over-evaluating themselves in both accentedness and comprehensibility.

The second set of analyses examined the relationships between L1 Mandarin speakers' self-ratings and the ratings assigned by listeners who also speak Mandarin as their L1. Pearson correlation tests (one-tailed) revealed moderate-strong associations between the speakers' self-ratings and the ratings assigned by L1 Mandarin listeners for accentedness, $r(39) = .64$, $p < .0001$, and moderate associations for comprehensibility, $r(39) = .54$, $p = .00013$. Results from paired samples T-test analyses indicated that the differences between self-ratings ($M = 5.96$, $SD = 1.75$) and L1 Mandarin listeners'

ratings (M = 5.88, SD = 1.46) were not significant for accentedness, p = .69, Cohen's d (effect size) = .054. The differences between self-ratings (M = 6.91, SD = 1.36) and L1 Mandarin listeners' ratings (M = 7.1, SD = 0.98) were, again, not significant for comprehensibility, p=.31, Cohen's d (effect size) = .16.

The third set of analyses examined the relationships between L1 Mandarin speakers' self-ratings and the ratings assigned by L2 English listeners who did not speak Mandarin as their L1. Pearson correlation tests (one-tailed) revealed moderate-strong associations between the speakers' self-ratings and the ratings assigned by L1 mixed listeners for accentedness, r(39) = .62, p <.0001, and moderate associations for comprehensibility, r(39) = .54, p = .00013. Results from paired samples T-test analyses indicated the differences between self-ratings (M = 5.96, SD = 1.75) and L1 mixed listeners' ratings (M = 5.05, SD = 1.5) were significant (p=.00023) for accentedness, Cohen's d (effect size) = .56. However, the differences between self-ratings (M = 6.91, SD = 1.36) and L1 mixed listeners' ratings (M=6.65, SD = 1.07) for comprehensibility were not significant, p=.17, Cohen's d (effect size) = .22.

As summarized in Table 1, the degree of agreement between L1 Mandarin speakers' pronunciation self-assessment and other listeners' assessment depends on the L1 background of the listeners. L1 Mandarin speakers' self-assessment was comparable to L1 Mandarin listeners' assessment in both comprehensibility and accentedness. L1 mixed listeners agreed with L1 Mandarin speakers in terms how comprehensible these L1 Mandarin speakers' speech sounded; however, as far as accentedness is concerned, L1 Mandarin speakers judged their own speech to be significantly less accented compared to

the assessment by L1 mixed listeners. L1 Mandarin speakers' self-assessment was significantly different from the assessment of L1 English listeners, with the L1 Mandarin speakers judging their own speech to be less accented and more comprehensible than the L1 English listeners did.

	Self vs. L1 English listeners	Self vs. L1 Mandarin listeners	Self vs. L1 mixed listeners
L1 Mandarin speakers (accentedness)	self-ratings (M = 5.96, SD = 1.75) L1 English listener ratings (M = 4.82, SD = 2.22) p<.005*	self-ratings (M = 5.96, SD = 1.75) L1 Mandarin listener ratings (M = 5.88, SD = 1.46) n.s., p>.5	self-ratings (M = 5.96, SD = 1.75) L1 mixed listener ratings (M = 5.05, SD = 1.5) p<.0005*
L1 Mandarin speakers (comprehensibility)	self-ratings (M = 6.91, SD = 1.36) L1 English listener ratings (M = 6.45, SD = 1.49) p<.05*	self-ratings (M = 6.91, SD = 1.36) L1 Mandarin listener ratings (M = 7.1, SD = 0.98) n.s., p>.3	self-ratings (M = 6.91, SD = 1.36) L1 mixed listener ratings (M=6.65, SD = 1.07) n.s., p>.1

Table 1. A comparison of self- vs. other-ratings, broken down by listener group, separately for accentedness and comprehensibility.

Discussion

The current study examined the associations between L1 Mandarin speakers' self-assessment of their L2 English pronunciation and the assessment of these speakers' pronunciation assigned by three different groups of listeners, L1 English listeners, L1 Mandarin listeners, and L1 mixed listeners. The results show that the degree of

agreement between self- and other-assessment depends on the L1 background of the listeners.

Use of Self-assessment in L2 Teaching, Learning, and Research

Considering the varied pronunciation demands in the diverse settings where English may be used in today's world, L2 learners are facing an increasing amount of responsibility to direct their own study in order to achieve their own individualized objectives. Self-assessment can enhance one's awareness of his/her own performance, and shift the decision-making process in the direction of the learner. In fact, self-assessment is included as a central component of DIALANG, a language diagnostic/placement test for 15 different European languages (see www.dialang.org). By providing learners with the opportunity to compare their self-assessment ratings with their actual performance in various skills areas, it is believed that the awareness of any potential discrepancies may lend insights into their language learning.

Self-assessment has been suggested as an alternative means of language assessment, as it enhances learners' awareness of the language learning process (Glover, 2011), promotes self-regulation and autonomy, and increases learner motivation (e.g., Noels, Clement, & Pelletier, 1999; Noels, Pelletier, Clement, & Vallerand, 2000). It has been found that in an autonomy-supportive environment, students were less likely to feel anxious in the learning process and less likely to give up L2 learning (Noels et al., 2000), which ultimately enhances students' achievement (Gardner, Tremblay & Masgoret, 1997; Horwitz et al., 1986; MacIntyre, Noels, & Clement, 1997). In sum, self-assessment is a very useful tool to help learners develop individualized learning goals, promote self-

regulation and self-efficacy, and achieve their objectives. It is of particular importance today as the goal of pronunciation learning is becoming increasingly varied and individualized.

However, despite its facilitative role in language leaning, self-assessment has been suggested to be used with caution due to its lack of validity and reliability. As far as the accuracy of self-assessment in L2 pronunciation is concerned, the current study revealed that while L2 learners' assessment of their own pronunciation does positively predict the assessments assigned by other listeners to some extent, their own assessment may or may not be comparable to that of other listeners, depending on the type of interlocutors intended. These results offer implications to the utilization of self-assessment in L2 classrooms. When the intended interlocutors are L1 English speakers, the findings of the present study suggest that there may be limited accuracy in pronunciation self-assessment, thus it may not be a reliable tool when accuracy is necessary and when the assessment is high-stakes. In comparison, in contexts when the interlocutors share the same L1 - Mandarin - with the speakers, pronunciation self-assessment may be a reliable measure of proficiency, and may be used as a complement to other traditional approaches for pronunciation assessment. In settings where the listeners are L2 English users from non-Mandarin L1 backgrounds, the reliability of L1 Mandarin speakers' self-assessment may depend on the construct under evaluation. Future studies that investigate additional L1s are needed to determine if the results of the current study may be generalizable to L2 English speakers from a non-Mandarin L1 background.

While it is useful for L2 educators to understand how self-assessment may align with other assessment differently depending on the contexts and interlocutors, it is also beneficial for L2 speakers themselves to be aware of the relative accuracy of their pronunciation self-assessment in different situations, which may be helpful for them to adjust their self-view and expectations.

The results from the current study also have implications for research design and policy-making. As Edele et al. (2015) pointed out, while studies that involve smaller samples have more liberty to apply language tests to assess L2 speakers' language skills, it is more realistic for population and household censuses, as well as large-scaled research studies, to measure language proficiency using self-assessment. In the past, a substantial proportion of research on immigrants' language proficiency, particularly in sociology and economics, has relied on data collected from self-reports (e.g., Berry et al., 2006; Carliner, 2000; Chiswick & Miller, 1995; Chiswick et al., 2004; Mouw & Xie, 1999; Pendakur & Pendakur, 2002; Van Tubergen & Kalmijn, 2009). The results of the current study suggest that pronunciation self-assessment may provide reliable measures of proficiency under certain circumstances.

The Alignment between Self- and Other-assessment

Shared evaluation criteria

Regarding the different levels of agreement observed here between self-assessment and the assessment of different types of listeners, one possible explanation taps into the field of social psychology. It has been suggested that people tend to have more accurate self-evaluations when the skills evaluated were specific rather than broad

and ambiguous (Zell & Krizan, 2014). Dunning, Meyerowitz, and Holzberg (1989) proposed that "faulty self-assessments occur because the meaning of most characteristics is ambiguous, which allows people to use self-serving trait definitions when providing self-evaluations" (p.1082). Their study found that people provide self-serving assessment to the extent that the trait is ambiguous, and that as the number of criteria increased, the subjects assessed themselves more accurately. The study also reported that the evidence and criteria that people use in self-evaluation is idiosyncratic, and requiring the subjects to evaluate themselves using a list generated by another individual led to more accurate self-appraisals. Since the majority of the L1 Mandarin subjects in the current study were born, raised, and have received their L2 English education in mainland China, there is reason to believe that these L1 Mandarin subjects may share among themselves similar concepts and expectations regarding English pronunciation compared to those subjects from a different cultural and educational background. In this sense, it is possible that the shared evaluation criteria contributed to the higher level of alignment between L1 Mandarin speakers' self-assessment and the assessment assigned by L1 Mandarin listeners.

Similarly, most of the L1 Mandarin speakers and L1 mixed speakers in the study were enrolled in ESL courses during the time the study was carried out. To a certain extent, their shared English learning experience could also have shaped some of their idiosyncratic view of English pronunciation, which could have impacted the assessment process, thus contributing to the degree of alignment between self- and other-assessment.

The impact of L1 background on L2 speech comprehension

While a shared set of criteria for pronunciation assessment generated from shared educational experience could have led to a higher degree of alignment between self- and other-assessment, it is also possible the patterns observed derived from a separate source - a possible advantage L2 listeners may have understanding L2 speech, which Bent and Bradlow (2003) have coined as the interlanguage speech intelligibility benefit (ISIB). The basic idea behind the ISIB is that speech intelligibility is enhanced between non-native interlocutors, compared to native/non-native interlocutors. An ISIB has been reported in situations when the L2 speakers and L2 listeners shared an L1 (matched ISIB) and did not have an L1 (mismatched ISIB) (Bent & Bradlow, 2003). In Hayes-Harb, Smith, Bent, and Bradlow (2008), the concept was further broken down into ISIB for listeners (ISIB-L) and ISIB for talkers (ISIB-T). ISIB for listeners refers to an advantage for L2 listeners over L1 listeners understanding L2 speech, while ISIB for talkers refers to an advantage for L2 speakers over L1 speakers when conversing with an L2 interlocutor. Both a matched benefit for listeners and a mismatched benefit and listeners could have contributed to the degree of alignment between self- and other-assessment observed here.

In the current study, while the comprehensibility ratings assigned by L1 English listeners were significantly lower than the speakers' comprehensibility self-assessment, the comprehensibility ratings assigned by the L1 Mandarin listeners and the L1 mixed listeners were comparable with speakers' self-assessment. These findings lend some support to a potential matched and mismatched benefit for listeners.

If L2 comprehension is indeed enhanced between L2-L2 interlocutors, compared

to L1-L2 interlocutors, the results here have implications for learners of a common L2, especially when the learners share the same L1. Though their impression of the comprehensibility of their own L2 speech may be on par with that perceived by their fellow English learners, these learners may have an inflated sense of how comprehensible their own and each other's speech sounds to listeners who are L1 users of English. This may need to be brought to the attention of L2 learners.

Conclusion and Limitations

In conclusion, the current study investigated the accuracy of L2 pronunciation self-assessment in relation to the assessment of different types of listeners. The study found that the degree of agreement between self- and other-assessment depends on the L1 background of the listeners and the constructs being evaluated. The L1 Mandarin listeners in the current study agreed with the L1 Mandarin speakers in terms of how comprehensible and accented these speakers' English speech sounded. L1 mixed listeners assigned ratings that were comparable to L1 Mandarin speakers' self-ratings in terms of how comprehensible these speakers' speech sounded, but the L1 mixed listeners judged these speakers' English to be significantly more accented than the speakers themselves did. L1 English listeners' assessment was significantly different from the L1 Mandarin speakers' self-assessment, with the L1 Mandarin speakers perceiving their own speech to be less accented and more comprehensible than the L1 English listeners did.

The results found here are consistent with the general belief that pronunciation difficulties are highly context-dependent - that listeners' L1 background significantly

affects their assessment of L2 speech (e.g. Jenkins, 2002; Stibbard & Lee, 2006). Therefore, the importance of taking contexts into consideration when making learning and pedagogical decisions is again brought to our attention.

While the results of the study may lend support to a potential interlanguage speech benefit for listeners in an L1 matched and mismatched situation, future studies focusing specifically on comparing the perceived comprehensibility of L1 and L2 listeners are needed to confirm if such an interlanguage speech comprehensibility benefit indeed can be found. Additionally, it has been suggested that the ISIB is likely mediated by factors such as properties of the speech itself (Munro et al., 2006), L2 proficiency of the listeners (e.g. Hayes-Harb et al., 2008; van Wijngaarden et al., 2002), L2 proficiency of the talkers (Bent & Bradlow, 2003; Hayes-Harb et al., 2008; van Wijngaarden, 2001; van Wijngaarden et al., 2002), and language environments (Xie & Fowler, 2013). Future studies are encouraged to take these factors into consideration in order to gain a fuller understanding of the impact L1 backgrounds have on L2 comprehension.

Additionally, it is unclear whether the results found here in regard to L2 comprehensibility may be generalized to L2 intelligibility. Future studies may explore intelligibility specifically to determine if the patterns observed here between self- and other-assessment hold when intelligibility is the construct under evaluation instead.

In the current study, L2 speakers of English from non-Mandarin L1 backgrounds were treated as a homogenous group (the L1 mixed group). Given that some languages may have more similarities in sound structures with Mandarin than other languages, it may be speculated that L2 English listeners from these different L1 backgrounds may

perceive the accentedness and comprehensibility of Mandarin-accented English differently, which may affect the alignment between self- and other-assessment. Future studies with carefully selected L1s are needed to provide a clearer picture in this respect.

Although it is not a goal of the current study to investigate the source of the differences in the assessment of accentedness vs. comprehensibility as well as in self- vs. other-assessment, it is likely that the constructs under evaluation and L1 backgrounds had an impact on the linguistic variables listeners attended to when assigning speech ratings. Trofimovich et al. (2016) found that while L2 speakers' self-ratings and L1 listeners' ratings of accentedness were based on the same linguistic factors (segmental, suprasegmental, and fluency), L2 speakers and L1 listeners took different linguistic dimensions into consideration when judging speech comprehensibility - L2 speakers' judgement of their own comprehensibility was linked only to segmental, suprasegmental, and fluency variables, whereas L1 English listeners' ratings were linked to additional factors such as lexicon, grammar, and discourse structure. Similarly, Foote (2015) found that different linguistic variables underlie the comprehensibility ratings of French-accented English speech assigned by L2 listeners from Mandarin, French, and Hindi L1 backgrounds. Future studies are encouraged to investigate how various linguistic factors may have differentially contributed to the perceived accentedness and comprehensibility by L1 listeners and L2 listeners from different L1 backgrounds. A better understanding in this direction will better inform L2 educators and L2 learners which linguistic dimensions are more facilitative to speech comprehension when conversing with different types of interlocutors.

CHAPTER FOUR

EXAMINATION OF AN INTERLANGUAGE SPEECH COMPREHENSBILITY BENEFIT

Introduction

Since the 1960s, comfortable intelligibility and comprehensibility have gradually replaced native-like accent, and become the tenet of pronunciation research and pedagogy (Derwing & Munro, 2005; Levis, 2005). However, while comprehensibility is an appropriate goal for L2 teaching and learning, it is also a very complex construct. Research has reported various factors that are associated with speech comprehensibility, such as rate of speech (e.g., Derwing & Munro, 2001), signal-to-noise ratio (e.g., van Wijngaarden, Steeneken , & Houtgast, 2002), whether talkers are speaking 'clearly' (e.g., Bradlow & Bent, 2002), word frequency (e.g., Bradlow & Pisoni, 1999), neighborhood density (e.g., Bradlow & Pisoni, 1999; Imai, Walley, & Flege, year), grammar (Ensz, 1982), vocabulary (Politzer, 1976), discourse (Albrechtsen, et al. 1980), and familiarity with the topic (Gass & Varonis, 1984). These pieces of evidence suggest that comprehensibility is not just a simple matter of the linguistic features of an utterance – listener factors also play an important role in speech comprehension.

Traditionally, L2 English pronunciation research and pedagogy have heavily relied on the perspectives of L1 English listeners. In today's world, the total number of L2 users of English has surpassed that of L1 speakers (Crystal, 2003), and the context in which English is used also varies drastically (e.g., ESL, EFL, EIL). As a result, an

increasing amount of communication is taking place between L2 users of English. Over the past 20 years, ESL pronunciation research has gradually moved toward investigating L2 pronunciation with L2-L2 interaction in mind (Jenkins, 2002; Walker, 2010). While some research has explored the role of L1 backgrounds in the judgment of L2 comprehensibility, much remains unknown. For instance, little is known about how L2 listeners may differ from L1 listeners in their judgement of L2 comprehensibility. Further, there is still much to be learned about how proficiency may impact how much difficulty listeners experience in L2 speech comprehension. A better understanding in these areas is needed to allow L2 educators and learners to make evidence-based pedagogical and learning decisions.

L2 Comprehensibility vs. Intelligibility

It has been proposed that both comprehensibility and intelligibility are important perceptual dimensions for speech evaluation (Derwing & Munro, 1997; Munro & Derwing, 1995, 1999). Intelligibility refers to "the extent to which a speaker's utterance is actually understood", which should be distinguished from comprehensibility, which refers to "the listener's estimation of difficulty in understanding an utterance" (Munro, Derwing, & Morton, 2006, p. 112). Comprehensibility and intelligibility are related but partially independent dimensions of L2 speech. For example, two utterances that are both "fully intelligible might entail perceptibly distinct degrees of processing difficulty, such that they are rated differently for comprehensibility" (p. 112). The two dimensions have also been assessed using different approaches. Comprehensibility is commonly rated using Likert-style rating scales, whereas intelligibility has been assessed using a number

of different methods, with the most common one being speech transcription. Between these two dimensions, comprehensibility may be the more important measure as far as successful communication is concerned. This is because "listeners can get frustrated talking with someone who requires a lot of effort to understand, even if those efforts are ultimately successful" (Foote, 2015, p. 38). Also, since intelligibility is commonly measured using speech transcription, even if a listener is able to recognize every word uttered, he/she may still struggle with the overall message conveyed.

Role of Language Background in L2 Comprehension

It has been long suggested that L2 speakers may be more intelligible to L2 listeners than L1 listeners. Back in 1969, Nash claimed that "A [non-native] speaker who cannot make himself understood when speaking English to a native English speaker will have no difficulty conversing in English with another [non-native] speaker" (p. 4). Similarly, Weinreich (1953) stated, "When the other interlocutor is also bilingual, the requirements of intelligibility... are drastically reduced" (p. 140).

Studies within language testing investigating if test takers have an advantage when sharing an L1 with the speakers of the listening materials have reported mixed results (Harding, 2012; Major, Fitzmaurice, Bunta, & Balasubramanian, 2002; Smith & Bisazza, 1983; Tauroza & Luk, 1997). For example, Major et al. (2002) had listeners from a variety of L1 backgrounds listen to TOEFL style lectures produced by speakers from different L1 backgrounds, and found that while L1 Spanish speakers performed significantly better when listening to other L1 Spanish speakers compared to Mandarin or Japanese-accented speakers, the same was not true for the Chinese L1 listeners, who

scored significantly lower when listening to Mandarin-accented English.

Outside of language testing, studies examining the impact of L1 backgrounds on L2 comprehension generally focus on the construct of intelligibility, and target the interlanguage speech intelligibility benefit (ISIB), a term coined by Bent and Bradlow (2003). The term "interlanguage" here refers to the language profile of a learner at some point in his or her L2 development (Selinker, 1972). The basic idea behind the ISIB is that speech intelligibility is enhanced between non-native interlocutors, compared to native/non-native interlocutors. It is important to note that the definition of the ISIB has gone through shifts over the years, during which the scope of "benefit" has, generally speaking, transitioned from the originally proposed definition (L2 speech being either equal to or more intelligible to L1 speech) (Bent & Bradlow, 2003) to a more literal definition that only includes the scenario when L2 listeners or speakers outperform L1 listeners or speakers (Stibbard & Lee, 2006). The more literal definition was adopted in the current study.

In Hayes-Harb, Smith, Bent, and Bradlow (2008), the concept was further broken down into ISIB for listeners (ISIB-L) and ISIB for talkers (ISIB-T). As discussed in Hayes-Harb et al. (2008), ISIB-T concerns cases when speech by L2 talkers is more intelligible to L2 listeners than speech by L1 talkers; in contrast, ISIB-L refers to cases where L2 speech is more intelligible to L2 listeners than it is to L1 listeners. In other words, ISIB-T compares the intelligibility of L1 vs. L2 talkers for L2 listeners, and ISIB-L compares the intelligibility of L2 talkers for L1 vs. L2 listeners. As summarized by the authors, the fundamental distinction between these two types of ISIB is "whether non-

native vs. native talkers are being compared (ISIB-T) or whether native vs. non-native listeners are being compared (ISIB-L)" (p. 665). The ISIB-T and ISIB-L have been found to be independent phenomena (Hayes-Harb et al., 2008; Xie & Fowler, 2013), between which the ISIB-L is of close relevance to the current study.

Depending on the L1 backgrounds of the L2 listeners and speakers, the ISIB was further broken down to matched ISIB and mismatched ISIB (Bent & Bradlow, 2003). The matched ISIB proposes that an L1 match between an L2 talker and L2 listener facilitates intelligibility, while the mismatched ISIB states that an L1 mismatch between an L2 talker and L2 listener facilitates intelligibility, both of which using L1-L2 communication as the base of comparison.

A few studies reported an ISIB-L in an L1 matched situation (Hayes-Harb et al., 2008; Imai et al., 2005; Munro et al., 2006; Xie & Fowler, 2013). In an attempt to investigate if L2 speakers' L1s affect their judgment of L2 speech, Munro, Derwing, and Morton (2006) found that Japanese-accented English was more intelligible to L1 Japanese listeners compared to L1 English listeners. Hayes-Harb et al. (2008) investigated the intelligibility of Mandarin-accented English for L1 English and L1 Mandarin listeners. Using a word identification task (minimal pairs that demonstrate word-final voicing contrast, such as "cub" and "cup"), the authors reported evidence for an ISIB-L, where the L1 Mandarin listeners were on average more accurate than the L1 English listeners at identifying words produced by L1 Mandarin speakers. Imai et al. (2003) compared the ability of L1 English listeners and L1 Spanish listeners at recognizing English words produced by an L1 Spanish speaker. The results revealed that

the L1 Spanish listeners outperformed L1 English listeners in the word recognition task for the words from dense lexical neighborhoods (i.e., words that have many similar sounding neighbors).

Only one study examined the ISIB-L in an L1 mismatched situation, and found no evidence for an ISIB-L (Munro et al. 2006). Munro et al. (2006) asked listeners from L1 Cantonese, Japanese, Mandarin, and English backgrounds to evaluate the same set of foreign-accented English speech from L1 speakers of Cantonese, Japanese, Polish, and Spanish. Regardless of the listeners' L1 backgrounds, different listener groups showed moderate to high correlations on their judgement of intelligibility, comprehensibility, and accentedness. The authors concluded that a mismatched ISIB-L effect was not found.

In sum, as demonstrated in Figure 7, studies that have investigated the ISIB-L found that when L2 listeners and L2 speakers shared an L1, L2 listeners perceived L2 speech to be more intelligible than L1 listeners did. When L2 listeners and speakers did not share an L1, the intelligibility of L2 speech was comparable as perceived by L2 listeners and L1 listeners.

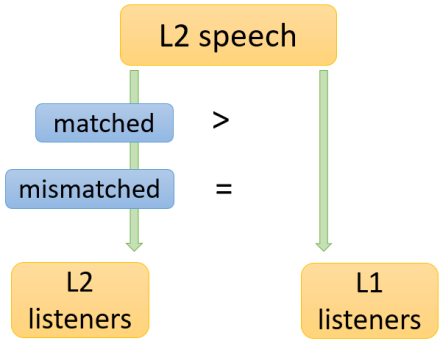

Figure 7. ISIB-L in L1 matched and mismatched situations.

Despite the supporting evidence provided by previous research, it has been suggested that while there can be an ISIB for L2 users, it is probably small and not consistently observable. For example, the matched ISIB-L observed in Munro et al. (2006) only held for L1 Japanese listeners and Japanese-accented English speech. No analogous benefit was found for L1 Cantonese listeners and L1 Cantonese talkers. The authors attributed the findings to the fact that "properties of the speech itself are a potent factor in determining how L2 speech is perceived, even when the listeners are from diverse language backgrounds" (p. 111). Inconsistent findings were also reported in Major et al. (2002) – while a matched ISIB was found with L1 Spanish listeners and Spanish-accented English speech, there was in fact an intelligibility disadvantage with L1 Chinese listeners and Chinese-accented English speech. Therefore, a general conclusion may be drawn that while there can be an ISIB, it is likely mediated by other factors.

One of these factors is the proficiency of L2 speakers and listeners, which has been reported to play an important mediating role in the ISIB. Bent and Bradlow (2003) investigated if there was an intelligibility benefit for L2 speech over L1 speech as judged by L2 listeners. They used speech samples collected via a sentence reading task from L1 speakers of Chinese, Korean, and English, and compared the intelligibility scores assigned by L1 Chinese, L1 Korean, L1 English listeners, as well as a group of listeners from mixed L1 backgrounds which exclude Chinese, Korean, and English. The results revealed an ISIB-T for high-proficiency L2 speakers, in both a matched and mismatched situation. Van Wijngaarden et al. (2002) reported that the listeners' L2 proficiency appeared to determine whether they find L1 or L2 talkers more intelligible. In their study,

L1 Dutch listeners who were more proficient in English than German demonstrated an ISIB-T for German but not for English. Specifically to the ISIB-L, the ISIB-L reported in Haye-Harb et al. (2008) held only for the low-proficiency listeners and low-proficiency speech.

As summarized in Figure 8, the only existing study which examined the impact of proficiency on ISIB-L reported an ISIB-L with low-proficiency L2 speakers and low-proficiency L2 listeners. No ISIB-L was detected with other proficiency pairs. So far, no study has examined the impact of proficiency on a potential ISIB-L in an L1 mismatched situation.

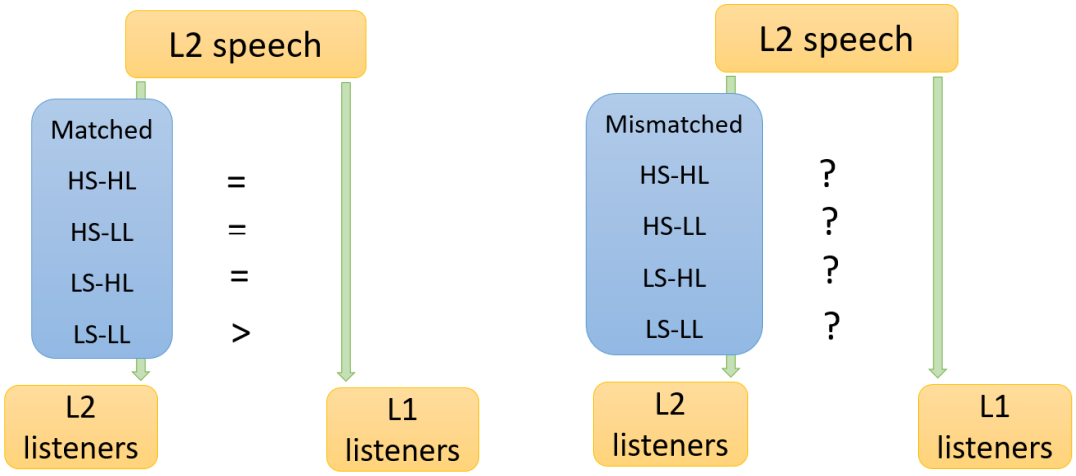

Figure 8. ISIB-L in L1 matched (left) and mismatched (right) situations when proficiency is taken into consideration (HS denotes high-proficiency speech, LS denotes low-proficiency speech, HL denotes high-proficiency listeners, LL denotes low-proficiency listeners).

Interlanguage Speech Benefit for Comprehensibility

Most of the studies targeting an interlanguage benefit for speech comprehension focused on intelligibility measures. However, to better understand communication in real-life settings, more information on the role of L1 backgrounds on L2 comprehensibility is

needed. One study that investigated comprehensibility is Munro et al. (2006). In terms of a comprehensibility benefit for talkers, the study found that the L1 Cantonese listeners found Cantonese-accented English speech to be easier to understand compared to the English speech of L1 Japanese, Polish, or Spanish speakers. The Japanese listeners found the Japanese-accented English speech easier to understand than the Cantonese-accented speech but not the speech by other L1 groups. In terms of a comprehensibility benefit for listeners, though not at a statistically significant level, L1 Japanese listeners judged the Japanese-accented English to be more comprehensible than L1 English listeners did, whereas the L1 Mandarin and L1 Cantonese listeners judged the L1 Japanese speech to be less comprehensible than the L1 English listeners did. Similarly, while the L1 Cantonese listeners judged the Cantonese-accented English to be more comprehensible than the L1 English listeners did, the L1 Mandarin and L1 Japanese listeners both judged the L1 Cantonese speech to be less comprehensible than the L1 English listeners did. Such findings lend some support to an interlanguage comprehensibility benefit for listeners in an L1 matched situation, but an interlanguage comprehensibility detriment for listeners in an L1 mismatched situation (Figure 9).

In another study by Foote (2015), listeners from Mandarin, French, Hindi, and English L1 backgrounds listened to English speech samples produced by speakers from Mandarin, French, and Hindi L1 backgrounds, and were asked to rate the speech samples for comprehensibility. With the analysis focusing on a potential benefit for talkers instead of listeners, the results revealed that the L1 Mandarin listeners rated the comprehensibility of Mandarin-accented English significantly higher than French- or

Hindi-accented English, whereas for the L1 French and Hindi listeners, language backgrounds did not significantly contribute to their assigned comprehensibility scores. Among the English speech samples from the speakers of the three L1 backgrounds, the ones by L1 Mandarin speakers were rated by L1 English listeners as the least proficient, which prompted the author to conclude that listeners who shared an L1 with lower-proficiency speakers may perceive L2 speech to be easier to understand than the speech of speakers from other L1 backgrounds.

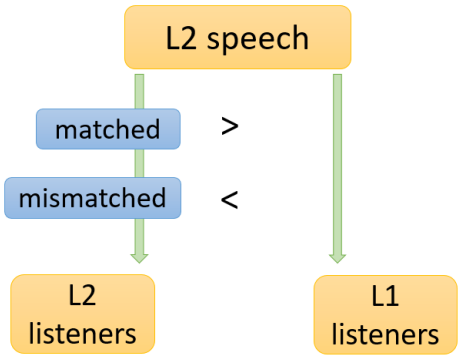

Figure 9. Interlanguage speech comprehensibility benefit in L1 matched and mismatched situations.

The Current Study

Existing studies investigating the role of language backgrounds in L2 comprehensibility are not only small in number, but also reported inconsistent findings. So far, only one study has examined a potential interlanguage speech benefit for comprehensibility for listeners, and no study has looked at the role proficiency may play in such a comprehensibility benefit. Given the importance of comprehensibility in successful communication, it is important to have a better understanding regarding if L2 listeners indeed enjoy a comprehensibility benefit listening to L2 speech, and whether

this benefit is mediated by factors such as listener and speaker proficiency level. Therefore, the chief goal of the current study was to examine whether listeners' L1 backgrounds affect how comprehensible L2 English speech was perceived to be, and if this potential benefit was limited to specific proficiency combinations of the speakers and listeners.

The following research questions were asked:

1) Do L2 English listeners who share the same L1 with the L2 English speakers perceive L2 speech to be significantly more comprehensible compared to L1 English listeners?
2) Do L2 English listeners who do not share an L1 with the L2 English speakers perceive L2 speech to be significantly more comprehensible compared to L1 English listeners?
3) Does English proficiency affect whether L2 English listeners perceive L2 English speech to be significantly more comprehensible than L1 English listeners when the L2 speakers and listeners share an L1?
4) Does English proficiency affect whether L2 English listeners perceive L2 English speech to be significantly more comprehensible than L1 English listeners when the L2 speakers and listeners do not share an L1?

In other words, the first two questions asked if there was a comprehensibility benefit in an L1 matched and mismatched situation, and the other two questions were interested in if the potential comprehensibility benefit was only limited to certain proficiency combinations of the L2 speakers and listeners.

Method

Participants

Speakers

The speakers were forty-one L1 Mandarin speakers (34 female, 7 male), with a mean age of 20.7 years (SD=3.7), who were enrolled in a university (39) or English language school (2) in the northeast U.S. during the time the study was carried out. Among those enrolled in a university, eleven were in graduate programs, twenty-eight in undergraduate programs. They had arrived in the U.S. to pursue studies at a mean age of 20 (SD=3.8), and have been studying in the U.S. for a mean of 8.6 months (SD=13.4). All speakers had taken either TOEFL iBT (39 subjects) or IELTS tests (2 subjects), which are high-stakes instruments that were used to assess the participants' ability to pursue university studies. The participants' mean overall scores were 100.3 (SD = 10.2) for TOEFL iBT, 25.6 (SD=4.0) for TOEFL listening, and 23.3 (SD=2.8) for TOEFL speaking. The TOEFL listening and speaking subscores were missing from one participant. The participants' mean overall scores were 6.75 (SD=1.1) for IELTS, 7.8 (SD=1.8) for IETLS listening, and 5.8 (SD=.4) for IETLS speaking. Among the TOEFL scores collected, four subjects had scores that were over 2 years old; the IELTS scores from both subjects were over 2 years old.

Excluding one participant, who accidentally left out the second page of the language background questionnaire, forty participants had studied English for an average of 12.9 years (SD = 2.8), primarily through formal instruction in primary, secondary, and university-level settings. The speakers self-rated their English ability at a mean of 5.6

(SD = 1.4) in speaking and 6.3 (SD = 1.6) in listening using 9-point Likert-type scales (1 = extremely poor, 9 = extremely fluent). Using 0–100% scales (0% = never, 100% = all the time), they also estimated their daily use of English at 57.5% (SD=22.7%).

Listeners

L1 English listeners

This group included eight inexperienced L1 English listeners (4 male, 4 female), with a mean age of 20.4 (SD = 3.7). All raters were L1 speakers of North American English, who reported using English an average of 98.6% (SD = 1.7%) of time in their daily life, out of which approximately 96.9% (SD=3.6%) of time was spent interacting with other L1 speakers of English (as opposed to L2 speakers). None of these subjects had studied Mandarin, and reported low familiarity with Mandarin-accented English (a self-rating that is equal to or smaller than 3 on a 1-9 scale, 1 – not at all familiar, 9 – very familiar). None of the subjects had any ESL/EFL experience or prior linguistic training experience. Given that all these subjects were residing in Boston during the time of the study, a city with a large international population, the exposure to and familiarity with accents other than Mandarin was considered acceptable. Accents that the subjects reported familiarity with to varying levels include German, French, Spanish, Italian, and Japanese.

L1 Mandarin listeners

This group included thirty-eight out of the forty-one participants in the speaker group. This is because two research meetings were scheduled for each participant in the speaker group (L1 Mandarin group). During the first meeting the L1 Mandarin

participants served as the speakers, and during the second meeting they served as listeners. Thirty-eight out of the original forty-one participants returned for the second research meeting.

L1 mixed listeners

The L1 mixed group were forty-one speakers (23 female,18 male) from an L1 background that is neither Mandarin nor English, with a mean age of 22.3 years (SD=4.3), who were enrolled in a university (15) or English language school (26) in the northeast U.S. during the time the study was carried out. Among those enrolled in a university, one was in a graduate program, fourteen in undergraduate programs. This group includes L1 speakers of Arabic (16), Creole (1), Korean (1), Thai (2), Kinyarwanda (1), Spanish (5), German (2), French (2), Polish (1), Swahili (1), Kalenjin (1), Italian (1), Kazakh (1), Portuguese (1), Greek (1), Japanese (6), and Vietnamese (1). One subject speaks both Arabic and French as L1s, one Swahili and Kalenjin as L1s, and one German and Polish. All three subjects were counted twice in the L1 counts. The participants had studied English for an average of 10.4 years (SD = 4.7), primarily through formal instruction in primary, secondary, and university-level settings. They arrived in the U.S. to pursue studies at a mean age of 21.7 (SD=4.2), and had been studying in the U.S. for a mean of 9.1 months (SD =11.4). Twenty-eight out of the forty-one subjects had taken either TOEFL iBT (18 subjects) or IELTS tests (10 subjects). The participants' mean overall scores were 92.1 (SD = 18.2) for TOEFL iBT, 24.8 (SD=4.5) for TOEFL listening, and 23.7 (SD=4.5) for TOEFL speaking. The TOEFL listening and speaking subscores were missing from two participants. The participants' mean overall

scores were 6.1 (SD=1.0) for IELTS, 6.2 (SD=1.3) for IETLS listening, and 6.4 (SD=.8) for IETLS speaking. Among the TOEFL scores, one subject had scores that were over 2 years old; none of the IELTS scores was over 2 years old. The speakers self-rated their English ability at a mean of 6.2 (SD = 1.3) in speaking and 6.4 (SD = 1.7) in listening using 9-point Likert-type scales (1 = extremely poor, 9 = extremely fluent). Using 0–100% scales (0% = never, 100% = all the time), they also estimated their daily use of English at 65.1% (SD=24.3%).

Materials and procedures

Speakers

During the first meeting with the L1 Mandarin group, each participant first filled out a questionnaire, which collected information such as age, gender, years of English learning, TOEFL iBT or IELTS score, years of residence in an English-medium country, age arriving in an English-speaking country, and etc. The participants were also instructed to submit a copy of their TOEFL or IETLS score report if available. In several cases, a score report was not available, and self-reported scores were accepted.

Afterward, each participant performed a picture narrative task. An eight-frame picture was used, which depicts two travelers bumping into each other and accidentally exchanging their identical suitcases (see Figure 10 below). Having first appeared in a study by Derwing, Munro & Thomson (2008), this picture has been used in a number of L2 pronunciation studies, and was selected here for the purpose of cross-study consistency and comparison.

Figure 10. Image used for speech elicitation.

The subjects were presented with the picture sequence and instructed to narrate a story of what happened in each image. There was no time limit imposed for preparation and narration. The narratives were recorded directly onto a computer and stored as digital audio files.

After this initial meeting, all the picture narrative recordings were subsequently normalized for volume by matching peak amplitude across files. They were also edited to remove all fillers and false starts at the beginning of the file and shortened to include only the initial 30 seconds of speech, consistent with prior research using 20-60 second samples to evaluate speech (e.g., Derwing et al., 2004).

Listeners

L1 Mandarin listeners

Thirty-eight out of the initial forty-one L1 Mandarin subjects returned several

months later and participated in the second research meeting, where they served as the L1 Mandarin listeners. During the second meeting, first they received a training session, which was followed by speech rating. The training session was to prepare the speakers to rate the comprehensibility of speech in the picture narrative task. The construct of comprehensibility was explained to the participants, and 3 practice speech files that are not relevant to the story depicted in the picture narrative task were subsequently be played for rating practice. The participants were invited to ask any question that they might have in regard to the construct. Afterward, the speakers used a 9-point scale to indicate how well they understood the concept (1 = don't understand at all, 9 = understand very well), and how comfortable they were using the construct to rate speech (1 = not comfortable at all, 9 = completely comfortable). Then, each participant listened to and rated the 40 speech samples produced by the L1 Mandarin speakers (41 minus the one produced by the subject him/herself) for comprehensibility. The participants were instructed that they could play and rate the recordings at their own pace, with an unlimited number of replays permitted, which is consistent with Trofimovich et al. (2016). The participants were also informed that although they were not required to listen to the entire recording to make a decision, they had to listen to at least 15 seconds of each recording, which is consistent with 15-30 second samples used to obtain listeners' impressionistic ratings of speech in prior research (e.g. Derwing et al., 2004). Additionally, the participants were also informed that all recordings were cut off after 30 seconds, and once they completed the ratings for a recording, they may not go back and change their assigned ratings.

L1 mixed listeners

First, each participant completed a language background questionnaire. Afterward, each participant performed the same picture narrative task as the one by L1 Mandarin speakers, following the same guidelines. Then the participants received the same training session, followed by speech rating. They rated the forty-one speech samples produced by all the L1 Mandarin speakers. The requirements and expectations of the speech rating procedure were the same as the L1 Mandarin listeners described above.

After all meetings with the L1 mixed listeners were completed, the picture narrative recordings produced by the L1 mixed listeners were edited following the same guidelines as the L1 Mandarin speech samples.

L1 English listeners

Each participant first completed a language background questionnaire. Afterward, they received a training session and proceeded to speech rating. The speech samples collected from the 41 Mandarin speakers were divided in two groups, 21 speech files in group 1 and 20 in group 2. Similarly, the speech samples from the 41 L1 mixed subjects were divided into two additional groups, 21 files in group 3 and 20 in group 4. Each L1 English listener rated the speech samples from one L1 Mandarin group and one L1 mixed group, which means each speech sample was rated by 4 randomly-chosen L1 English listeners. Different from the L1 Mandarin listeners and the L1 mixed listeners, the L1 English listeners rated both the comprehensibility and the accentedness of each speech sample for reasons that will be discussed later in the analysis section.

For speech rating, a 9-point Likert-type numerical scale was chosen as the instrument for speech measurement (see Figure 11).

Accentedness: This refers to how much a speaker's speech is influenced by his/her native language and/or is colored by other non-native features.								
1 heavily accented	2	3	4	5	6	7	8	9 not accented at all

Comprehensibility: This refers to how much effort it takes to understand what someone is saying. If you can understand with ease, then a speaker is highly comprehensible. However, if you struggle and must listen very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility.								
1 hard to understand	2	3	4	5	6	7	8	9 easy to understand

(adapted from Saito & Shintani, 2016)

Figure 11. 9-Point scale used for speech rating.

Analysis and Results

Cronbach's alpha, a measure of inter-rater reliability, was computed across the English listeners' ratings, separately for accent and comprehensibility. The obtained coefficients were 0.95 for accent and 0.89 for comprehensibility, exceeding the benchmark value of .70 - .80 (Larson-Hall, 2010). Fleiss' κ was run to determine if there was agreement within the L1 English listeners' judgement in terms of how accented and comprehensible each speech sample sounds. At an agreement window of ±1, κ (accent) = 48.2% and κ (comprehensibility) = 38.0%. At an agreement window of ±2, κ (accent) =

69.1% and κ (comprehensibility) = 61.5%. The Fleiss' κ values, taken together with the Cronbach's alphas, provide strong evidence that there is coarse inter-rater reliability. Therefore, a single accentedness and comprehensibility score was derived for each speaker by averaging across the ratings assigned by the four randomly-chosen L1 English listeners.

Cronbach's alpha was also computed across the ratings assigned for each Mandarin-accented speech sample by all L1 Mandarin listeners and by all L1 mixed listeners for comprehensibility. For the L1 Mandarin listeners' ratings, the obtained coefficients was 0.97; and for the L1 mixed listeners' ratings, the obtained coefficient was 0.97. Since the alphas all exceeded the benchmark value of .70 - .80 (Larson-Hall, 2010), a single comprehensibility score was derived for each speaker by averaging across all ratings assigned by the L1 Mandarin listeners, and across all ratings assigned by the L1 mixed listeners.

The first set of analyses investigated if there were any significant differences between the comprehensibility ratings of the Mandarin-accented English speech assigned by L1 Mandarin listeners and L1 English listeners. Results from paired T-test analyses indicated that L1 Mandarin listeners (M = 7.1, SD = .98) perceived the Mandarin-accented speech to be significantly more comprehensible than L1 English listeners (M = 6.45, SD = 1.49), $p<.0001$, Cohen's d (effect size) = .53.

The second set of analyses investigated if there were any significant differences between the comprehensibility ratings of the Mandarin-accented English speech assigned by L1 mixed listeners and L1 English listeners. Results from paired T-test analyses

showed no significant difference between the comprehensibility ratings assigned by L1 mixed listeners (M = 6.65, SD = 1.07) and L1 English listeners (M = 6.45, SD = 1.49), p=.10, Cohen d=.16.

The third set of analyses investigated if the proficiency level of the L1 Mandarin speakers and L1 Mandarin listeners had any impact on the comprehensibility benefit unveiled in the first set of analyses. To conduct this analysis, first the L1 Mandarin subjects were categorized into high- and low-proficiency groups. In existing studies examining the impact of proficiency on the perception of L2 speech, proficiency has been measured using different approaches. It has been operationalized as accentedness (Hayes-Harb et al., 2008; Imai et al., 2005; Stibbard & Lee, 2006), intelligibility (Bent & Bradlow, 2003), listening proficiency (Xie & Fowler, 2013), or measured through self-reports (van Wijngaarden et al., 2002). In the current study, similar to Haye-Harb et al. (2008), Imai et al. (2005) and Stibbard and Lee (2006), proficiency, or more specifically, phonological proficiency, was operationalized as accentedness, where those with lower accentedness ratings judged by L1 English listeners were considered more proficient in the L2.

As shown in Figure 12, the accentedness of L1 Mandarin speakers, as rated by L1 English listeners, ranges from 1.25 to 8.75, with a mean of 4.82 and the standard deviation of 2.22. High proficiency was defined as an accentedness rating of 6.5-9, while low proficiency a rating between 1-3.5. This yielded 12 high-proficiency L1 Mandarin speakers (M=7.63, SD=0.68), and 13 low-proficiency L1 Mandarin speakers (M=2.33, SD=0.84).

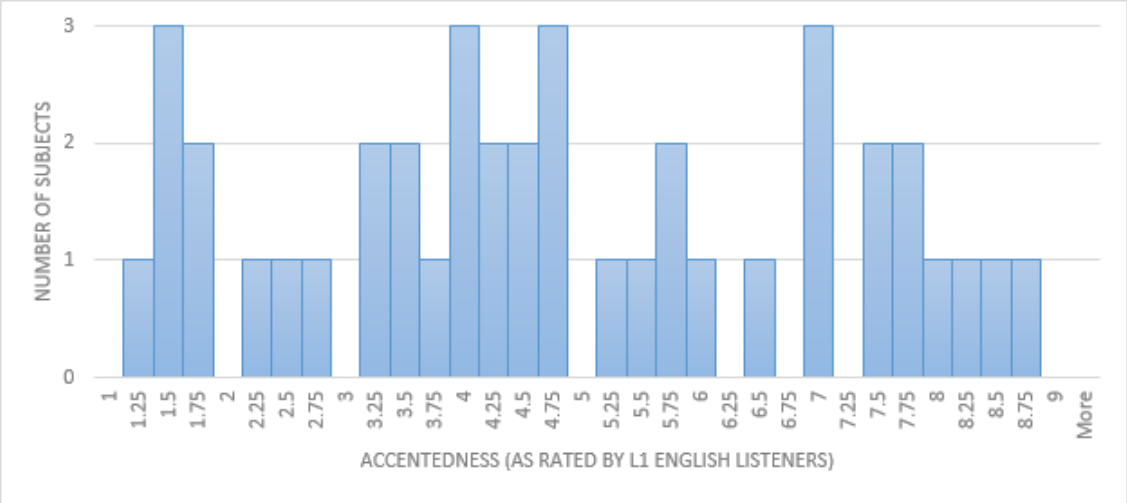

Figure 12. Distribution of L1 Mandarin speakers' phonological proficiency as measured by accentedness scores assigned by L1 English listeners.

For each Mandarin-accented speech sample in the low- and high-proficiency group, an average score of L1 English listeners' comprehensibility ratings was calculated. The same was done with L1 Mandarin listeners' comprehensibility ratings as discussed below.

As discussed earlier, in the current study the L1 Mandarin speakers returned for a second research meeting, where they served as the L1 Mandarin listeners. Thirty-eight out of the initial forty-one participants returned for the second meeting. All twelve high-proficiency L1 Mandarin speakers returned for the second research meeting, and rated the speech samples of all the other Mandarin L1 speakers. In comparison, eleven out of the thirteen low-proficiency L1 Mandarin speakers returned for the second meeting, and rated the speech by the other L1 Mandarin speakers. Therefore, each high-proficiency L1 Mandarin speech sample (n= 12) was rated by eleven high-proficiency L1 Mandarin listeners and eleven low-proficiency L1 Mandarin listeners; each low-proficiency L1

Mandarin speech sample (n=13) was rated by twelve high-proficiency L1 Mandarin listeners and ten low-proficiency L1 Mandarin listeners. For each speaker within the high-proficiency and low-proficiency group, an average score was calculated out of the scores assigned by high-proficiency listeners and low-proficiency listeners for comprehensibility.

For the high-proficiency L1 Mandarin speech (HP-M speech), paired T-test analyses were carried out to test if there were any significant differences between the comprehensibility ratings assigned by L1 English listeners (NE listeners) and high-proficiency L1 Mandarin listeners (HP-M listeners), and between the comprehensibility ratings assigned by L1 English listeners and low-proficiency L1 Mandarin listeners (LP-M listeners). Results revealed that HP-M listeners (M = 8.36, SD = .46) perceived the HP-M speech to be significantly more comprehensible than the NE listeners did (M = 7.90, SD = .79), p=.005, Cohen d=.75. There was no significant difference between the comprehensibility ratings assigned by the LP-M listeners (M = 7.92, SD = .64) and NE listeners (M = 7.90, SD = .79), p=.89, Cohen d=.03.

For the low-proficiency L1 Mandarin speech (LP-M speech), paired T-test analyses were carried out to test if there were any significant differences between the comprehensibility ratings assigned by NE listeners and high-proficiency L1 Mandarin listeners (HP-M listeners), and between the comprehensibility ratings assigned by L1 English listeners and low-proficiency L1 Mandarin listeners (LP-M listeners). Results revealed that HP-M listeners perceived the LP-M speech to be significantly more comprehensible (M = 6.68, SD = .84) than the NE listeners did (M = 5.5, SD = 1.1),

p=.00015, Cohen d=1.21. The LP-M listeners also perceived the LP-M speech to be more comprehensible (M = 6.18, SD = 1.17) than the NE listeners did (M = 5.5, SD = 1.1), p=.0054, Cohen d=.59.

	HP-M listeners vs. NE listeners	LP-M listeners vs. NE listeners
HP-M speech	HP-M listeners judged the speech to be more comprehensible than did the NE listeners ($p<.05$)	n.s., $p>.5$
LP-M speech	HP-M listeners judged the speech to be more comprehensible than did the NE listeners ($p<.005$)	LP-M listeners judged the speech to be more comprehensible than did the NE listeners ($p<.05$)

Table 2. A comparison of comprehensibility ratings, broken down by listener and speaker group.

The fourth set of analyses investigated if the proficiency level of the L1 Mandarin speakers and L1 mixed listeners had any impact on the existence of an interlanguage comprehensibility benefit. To conduct this analysis, the L1 mixed subjects were first categorized into high- and low-proficiency groups. As shown Figure 13, the accentedness of L1 mixed speakers, as rated by English L1 listeners, ranges from 2 to 9, with a mean of 5.59 and the standard deviation of 1.87. Following the same guideline that high proficiency covers an accentedness rating of 6.5-9, and low proficiency 1-3.5, sixteen L1 mixed speakers fell into the high-proficiency group (M=7.55, SD=0.71), and six in the low-proficiency group (M=2.75, SD=0.52).

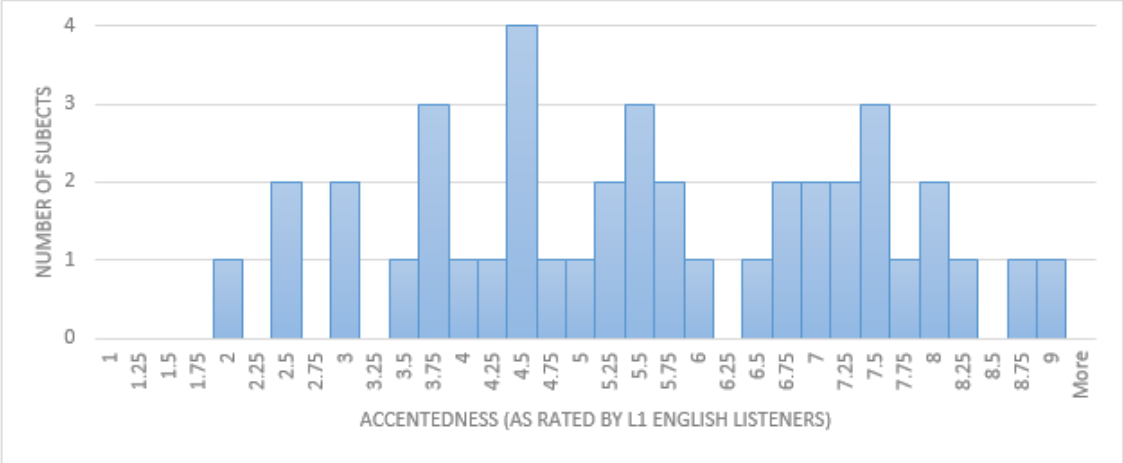

Figure 13. Distribution of mixed L1 subjects' phonological proficiency as measured by accentedness scores assigned by L1 English listeners.

For each L1 Mandarin speech sample within the high-proficiency and low-proficiency group, an average score was calculated out of the scores assigned by the sixteen high-proficiency L1 mixed listeners and six low-proficiency L1 mixed listeners for comprehensibility.

For the high-proficiency L1 Mandarin speech (HP-M speech), paired t-test analyses were carried out to test if there were any significant differences between the comprehensibility ratings assigned by L1 English listeners (NE listeners) and high-proficiency L1 mixed listeners (HP-X listeners), and between the comprehensibility ratings assigned by L1 English listeners and low-proficiency L1 mixed listeners (LP-X listeners). Results revealed that there was no significant difference between the comprehensibility ratings assigned by HP-X listeners (M = 7.66, SD = .64) and NE listeners (M = 7.90, SD = .79), p=.058, Cohen d=.33. In comparison, the differences between the comprehensibility ratings assigned by LP-X listeners and NE listeners were

significant, with the NE listeners perceiving the speech to be more comprehensible (M = 7.90, SD = .79) than the LP-X listeners did (M = 7.25, SD = .81), p=.033, Cohen d=.81.

For the low-proficiency L1 Mandarin speech (LP-M speech), paired t-test analyses were carried out to test if there were any significant differences between the comprehensibility ratings assigned by NE listeners and high-proficiency L1 mixed listeners (HP-X listeners), and between the comprehensibility ratings assigned by L1 English listeners and low-proficiency L1 mixed listeners (LP-X listeners). Results revealed that there was no significant difference between the comprehensibility ratings assigned by the HP-X listeners (M = 5.11, SD = 1.22) and the L1 English listeners (M = 5.5, SD = 1.1), p=.079, Cohen d=.34. However, the differences between the comprehensibility ratings assigned by LP-X listeners and L1 English listeners were significant, with the LP-X listeners perceiving the speech to be more comprehensible (M = 6.23, SD = .96) than the L1 English listeners did (M = 5.5, SD = 1.1), p=.018, Cohen d=.71.

	HP-X listeners vs. NE listeners	LP-X listeners vs. NE listeners
HP-M speech	n.s., $p>.05$	LP-X listeners judged the speech to be less comprehensible than did the NE listeners ($p<.05$)
LP-M speech	n.s., $p>.05$	LP-X listeners judged the speech to be more comprehensible than did the NE listeners ($p<.05$)

Table 3. A comparison of comprehensibility ratings, broken down by listener and speaker group.

Discussion

The present study examined if there was an interlanguage speech comprehensibility benefit for listeners in both an L1 matched and mismatched situation, and if the proficiency level of L2 speakers and listeners affected the presence or absence of a potential comprehensibility benefit.

In an L1 matched situation, a comparison between the comprehensibility ratings of Mandarin-accented English speech assigned by L1 Mandarin listeners and L1 English listeners revealed an interlanguage speech comprehensibility benefit. In other words, L1 Mandarin listeners perceived the Mandarin-accented speech to be significantly more comprehensible than did the L1 English listeners. When proficiency was taken into consideration, a comprehensibility benefit was again observed with LP-M speakers/LP-M listeners, LP-M speakers/HP-M listeners, and HP-M speakers/HP-M listeners. The results here are consistent with Hayes-Harb et al. (2008) in that LP-M listeners in the current study had an advantage over L1 English listeners at comprehending LP Mandarin-accented speech, although intelligibility, instead of comprehensibility, was the measure used in Hayes-Harb et al. (2008). Different from Hayes-Harb et al. (2008), the current study also found a comprehensibility benefit with two additional proficiency combinations: HP-M speakers/HP-M listeners, and LP-M speakers/HP-M listeners. The only pair for which no interlanguage comprehensibility benefit was observed was the HP-M speakers/LP-M listeners - the LP-M listeners judged HP-M speech to be equally comprehensible as the L1 English listeners did.

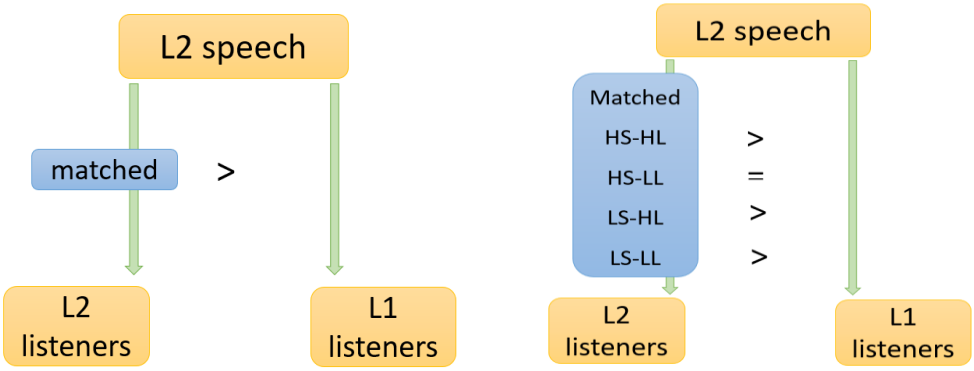

Figure 14. A comparison of the comprehensibility ratings assigned by L1 Mandarin listeners and L1 English listeners (left – overall, right – proficiency considered).

The interlanguage speech comprehensibility benefit observed here is in line with the large body of research indicating that L2 speech is likely to be more easily understandable to other L2 speakers from the same L1 background. From the perspective of cross-language speech perception, such a shared-L1 advantage is based on the principle that "L2 accents are primarily characterized by transfer from the L1", therefore when a listener shares an L1 with the speaker, he or she will have "an intimate familiarity with the phonological patterns of that speaker's L2 accent" (Harding, 2011, p. 165). According to Bent and Bradlow (2003), the overall shared phonetic and phonological knowledge between L2 talkers and L2 listeners from the same L1 background is likely to be more extensive than an L1/L2 pair. Thus, when an L1 is shared, an L2 listener is more likely better equipped to interpret certain acoustic–phonetic features of L2 speech as the talker intended them to be interpreted, even though they may "deviate markedly from the target language norm" (Bent & Bradlow, 2003, p. 1607).

Although the LP-M listeners did not judge the HP-M speech to be significantly more comprehensible than L1 English listeners did, the ratings by the LP-M listeners

were also not significantly lower than those assigned by the L1 English listeners. This itself may be interpreted as additional supporting evidence for a matched comprehensibility benefit. Since these listeners are low-proficiency English users with limited experience hearing English speech, it is to be expected that L1 English users should outperform these people at comprehending English speech. However, as the results revealed, the L1 English listeners experienced a similar level of difficulty understanding the HP-M speech as these low-proficiency L2 users, which indicates that these L2 English listeners' lack of experience with English was offset somehow, possibly by sharing an L1 with the speakers.

The current study found a matched interlanguage benefit for three different proficiency combinations, whereas this interlanguage benefit was only observed with low-proficiency listeners and low-proficiency speakers in Hayes-Harb et al. (2008). There are several possible explanations for the different patterns observed in these two studies. One of these explanations is the different constructs measured and the different methods adopted to measure these constructs. In Hayes-Harb et al. (2008), intelligibility was the construct under evaluation, which was identified as word identification accuracy in a word identification task. In their study, the subjects listened to isolated target word tokens in a minimal pair fashion, and identified each word they heard by selecting the written word that matched the auditory stimulus (e.g., hear 'cub'; identify as 'cub' or 'cup'). In the current study, comprehensibility was tested rather than intelligibility, and was measured by assigning a comprehensibility score (on a scale of 1-9) after hearing a story narrative. It is difficult to determine at which stage of spoken language processing

the interlanguage benefit arose in the current study, as perceiving speech requires processing on many different levels and the comprehensibility measure used here was an off-line measure that reflected the accumulation of processing on multiple levels. In comparison, in Hayes-Harb et al. (2008), since a predetermined set of words was used, it may be assumed that the benefit uncovered derived from the bottom-up processing of the auditory signal instead of differences in lexical choices, syntactic structures, or sociolinguistic factors. Therefore, it is likely that the interlanguage benefit observed in the two studies came from different sources. Additional tests that specifically tap into various levels of speech processing are needed to determine speech processing at which levels were at play in the interlanguage comprehensibility benefit observed here. Another possible source of the variations in results between the two studies is the differences in the proficiency of the talkers and listeners. While both Hayes-Harb et al. (2008) and the current study used accentedness as the criteria to rank proficiency, Hayes-Harb et al. (2008) categorized the lowest and highest thirds within their data sample as the low-proficiency and high-proficiency group, whereas the current study identified an accentedness score within 6.5-9 to be high proficiency, and 1-3.5 to be low proficiency. Additional studies that specifically compare different proficiency ranges are needed to determine if the variations in proficiency was indeed a source of the differences observed here.

In an L1 mismatched situation, a comparison between the comprehensibility ratings of Mandarin-accented English speech assigned by L1 mixed listeners and L1 English listeners did not reveal a mismatched benefit for comprehensibility. In other

words, L1 mixed listeners did not perceive the Mandarin-accented English speech to be more comprehensible than the L1 English listeners did. In fact, overall speaking, the comprehensibility ratings of Mandarin-accented speech assigned by L1 mixed listeners and L1 English listeners were on par with each other. When proficiency was taken into consideration, the picture was more complex. An interlanguage comprehensibility benefit was observed with LP-M speakers and LP-X listeners, whereas an interlanguage comprehensibility detriment was observed with HP-M speakers and LP-X listeners. With the other two proficiency combinations – LP-M speakers/HP-X listeners and HP-M speakers/HP-X listeners, L1 mixed listeners' comprehensibility ratings were comparable to L1 English listeners'.

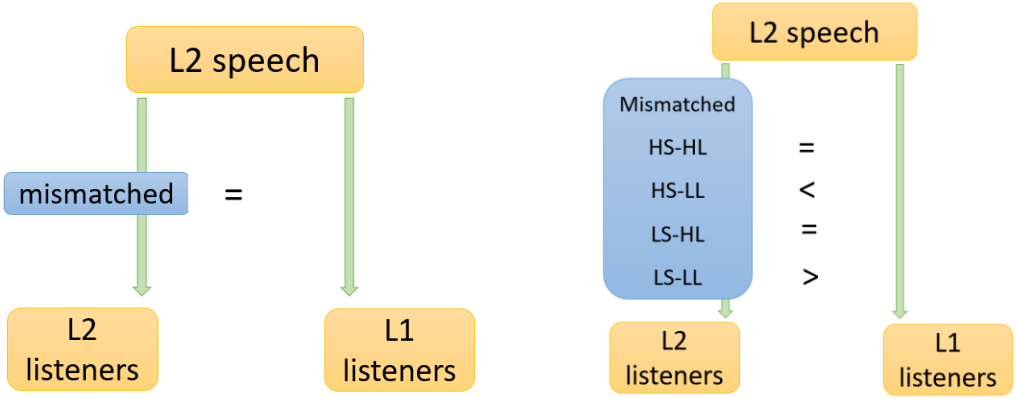

Figure 15. A comparison of the comprehensibility ratings assigned by L1 mixed listeners and L1 English listeners (left – overall, right – proficiency considered).

Compared to the L1 matched situation, a comprehensibility benefit was not consistently observed in an L1 mismatched situation. This is not surprising as the overall shared phonetic and phonological knowledge between L2 users who are from different L1 backgrounds is likely less extensive than that between those who share an L1.

Therefore, in an L1 mismatched situation, the listeners are less likely to be able to interpret certain L1-influenced acoustic–phonetic features of L2 speech as well as they would in an L1 matched situation.

In terms of why the LP-X listeners perceived the LP-M speech to be more comprehensible than did the L1 English listeners, a possible explanation is the speakers' and listeners' shared knowledge of the structure of the target language, as well as the influence of general strategies that listeners and speakers adopt when learning to produce and comprehend a foreign language (Bent & Bradlow, 2003). One example Bent and Bradlow (2003) provided is the production and perception of word-final stop consonants in American English. Since these consonants are typically unreleased, L1 listeners typically rely heavily on cues in other parts of the utterance to identify these consonants, which L2 listeners are often unable to do. In their own L2 English speech, since these L2 speakers have not yet mastered all the details of American English allophony, they may produce particularly salient word-final stop consonant releases, which in fact facilitates the comprehension of their speech for other L2 listeners. Such a benefit may be particularly pronounced among low-proficiency L2 users as "learners at a lower level of proficiency are more similar to each other in the nature of their L2 phonological representations and/or the ways in which they phonetically implement L2 phonological contrasts than are learners at higher proficiencies" (Hayes-Harb et al., 2008). As learners reach higher levels of L2 proficiency, they are likely to have more varied experience with the target language, thus "exhibit more diversity in their phonological systems" (p. 675).

In terms of the interlanguage comprehensibility detriment observed with HP-M

speakers and LP-X listeners, it is to be expected for low-proficiency L2 users to perform less well compared to L1 English listeners at comprehending high-proficiency English speech. The fact that what is observed here was not observed with low-proficiency L1 Mandarin listeners further lends support to a matched comprehensibility benefit. While both low-proficiency groups had limited experience hearing English speech, the L1 Mandarin listeners judged the speech to be significantly more comprehensible than the L1 mixed listeners did. The fact that this pattern was observed despite the overall higher proficiency of the LP-X group (M=2.75) compared to the LP-M group (M=2.33) further reinstates that when L2 users share an L1, the effort it takes for them to understand each other's L2 is significantly reduced.

Limitations

There are a number of limitations to the current study that should be addressed in future research. First, it is unclear if the findings of the current study will hold when other types of speech are utilized. It has been reported that the type of speech and the cognitive demand involved in the elicitation task affect L2 speech comprehension. Given that the current study only utilized picture narratives, future studies that include and compare different speech types would further expand our understanding in this area.

Secondly, both the current study and Hayes-Harb et al. (2008) used L1 Mandarin speakers to test a matched interlanguage speech benefit for listeners. It is conceivable that there is some sociolinguistic factor at play which may have contributed to the findings. Further research targeting other L1s is needed to test if the findings here are generalizable

to speakers and listeners of other L1s.

Additionally, the current study only examined the impact of two factors on L2 comprehensibility, L1 background and proficiency. It is quite possible that there are additional factors that also have an impact on L2 comprehension, such as properties of speech itself, including segmental sounds, prosody, grammatical accuracy, vocabulary use, syntactic structures, and L1-based discourse structure, or factors that are unrelated to speech, listener bias and expectations for example. As noted by Foote (2015), "the entire concept of speech comprehensibility is multifaceted, comprised of qualities of speech and characteristics of the listener, not to mention many variables relating to the context of a given interaction" (p. 60). Future studies should take into consideration these additional factors, and examine if and how they may interact with L1 backgrounds and proficiency, and affect the comprehensibility benefit observed.

Moreover, an issue that was raised in previous ISIB studies is whether mismatched ISIB is a separate phenomenon from matched ISIB (Bent & Bradlow, 2003). Bent and Bradlow (2003) reported a mismatched ISIB-T using L1 Korean and L1 Chinese subjects, and speculated that the two groups' shared phonological similarities and geographical and cultural proximity could have contributed to the mismatched ISIB, which makes the mismatched ISIB observed "just another manifestation of the matched interlanguage benefit rather than a separate phenomenon" (p. 1607). In response to this speculation, Stibbard and Lee (2006) investigated an ISIB-T using Saudi Arabian and Korean speakers, two L1 groups that did not share as much phonological and cultural similarity, and found no evidence for a mismatched ISIB-T. The L1 mixed group in the

current study includes a variety of L1s, some of which have more similarities with Mandarin in terms of phonological structures and culture than others. Due to the limited number of subjects from each different L1 included in the current study, the data here do not provide enough evidence in terms of if the mismatched benefit observed was a separate phenomenon from the matched benefit. Future studies with carefully selected languages are needed in order to take a closer look at this specific question.

Conclusion

Consistent with previous studies reporting an ISIB-L pattern (Hayes-Harb et al., 2008; Imai et al., 2005; Munro et al., 2006; Weinreich, 1953; Xie & Fowler, 2013), the current study found evidence supporting a matched interlanguage speech benefit for listeners when comprehensibility is taken into consideration. The term "benefit" is used here to mean only cases in which higher comprehensibility scores were given by the listeners, in contrast to the Bent and Bradlow (2003) use of the term to include equal scores. When speaker and listener proficiency was taken into consideration, such a matched comprehensibility benefit for listeners was found in 3 out of 4 proficiency combinations (HP-M speakers/HP-M listeners, LP-M speakers/HP-M listeners, LP-M speakers/LP-M listeners). In terms of a mismatched interlanguage speech comprehensibility benefit for listeners, although no evidence was found to support an overall mismatched benefit, when the proficiency of speakers and listeners was taken into consideration, a mismatched benefit was detected with the LP-M speakers/LP-X listeners, but a mismatched comprehensibility detriment was detected with the HP-M speaker/LP-

X listeners.

With globalization and English becoming a lingua franca, it is increasingly common for L2 English users to communicate with each other in English. The observed matched interlanguage speech benefit for comprehensibility, in combination with other sociolinguistic factors, will likely encourage the establishment of new pronunciation norms within different communities of L2 users of English. These communities could be on a scale of an entire region, such as southeast Asia (Deterding & Kirkpatrick, 2006), specific countries, such as India or Singapore, or science labs and ESL classrooms in English-medium universities. While L2 English speakers who are in the beginning stage of English learning or those who share an L1 with the listeners may be able to adjust their production or use their shared knowledge base to compensate for their limited experience with English, the success they experience within their own communities may not extend to the bigger environment, when they communicate with L1 English speakers or more proficient L2 English speakers from different L1 backgrounds. Thus, it is important for L2 educators and learners to recognize their potential false sense of L2 comprehensibility, approach L2 pronunciation teaching and learning with specific communicative contexts in mind, and learn about the different expectations and demands associated with English use in various contexts and with different interlocutors.

While it is important for educators to encourage learners to interact with different types of interlocutors and help learners be understood in these different situations, the findings of this study also highlight the importance for educators to train learners to understand different types of potential interlocutors. Given that experience with accented

speech is associated with improved comprehension (e.g. Kennedy & Trofimovich, 2008; Winke, Gass, & Myford, 2013), it may be critical for educators to integrate speech models by speakers from different L1 backgrounds in their English classrooms in order to prepare their students to become truly competent English users in today's world.

CHAPTER FIVE

CONCLUSION

Although an increasing amount of research has been done in the area of L2 pronunciation assessment in the past few decades, much remains unknown in terms of how L2 speakers' pronunciation self-assessment compares to the assessment of these speakers' pronunciation judged by their interlocutors, particularly considering the various interactional contexts and the different types of interlocutors that may be involved. This dissertation presents an attempt to gain a better understanding in this particular area. Each of the 3 studies in this dissertation had its own specific objectives, though together, they intended to address two primary objectives: 1) to increase our understanding in terms of how L2 pronunciation self-assessment compares to other-assessment in different interactional contexts, and 2) to understand why the level of agreement between self- and other-assessment may differ when different types of listeners are taken into consideration. In this chapter, I will give a brief summary of these three studies, connecting each one, and review the key findings from each. Then I draw general conclusions from the studies and discuss the general implications.

Overview of Key Findings

Study 1 and Study 2 intended to address the first objective of this dissertation: to understand how L2 pronunciation self-assessment compares to other-assessment in different interactional contexts. These studies were motivated by the idea that having an accurate assessment of one's own pronunciation is important to their success in English

acquisition and their social or professional life. The specific objective of Study 1 was to determine if L2 English speakers were able to assess their own English pronunciation accurately in relation to the judgment by L1 English listeners. Eighty-two L2 speakers of English from a variety of L1 backgrounds completed a picture narrative task and rated the accentedness and comprehensibility of their own speech. The recordings of their speech were later rated by eight L1 English listeners for accentedness and comprehensibility. The self- and other-ratings were compared using correlation and paired t-test analyses. The key findings from Study 1 were: 1) although L2 speakers' self-assessment of comprehensibility and accentedness were moderately associated with the assessment by L1 English listeners, the ratings assigned by these two groups (self vs. L1 English listener) were significantly different from each other; and 2) speakers at the bottom of the accentedness and comprehensibility scales overestimated their performance while speakers at the top of each scale underestimated it.

Given that there are more L2 users of English than L1 users in today's world, L2 English speakers may frequently find themselves using English to communicate with other L2 speakers of English. With this complexity of interaction in mind, Study 2 expanded the scope of interlocutors to include not only L1 English listeners, but also listeners who are L2 users of English. Additionally, given that the results from Study 1 suggested that the speakers' L1 backgrounds may be a source of differences in how well self-assessment aligned with other-assessment, Study 2 focused on the largest cohort in the sample, L1 Mandarin speakers, and compared L1 Mandarin speakers' pronunciation self-ratings to the ratings assigned by L1 English listeners, L1 Mandarin listeners, and

listeners whose L1 is neither English nor Mandarin. The primary objective of Study 2 was to expand on the findings of Study 1 and further investigate the alignment between pronunciation self- and other-assessment when different types of listeners were taken into consideration. Forty-one L1 Mandarin speakers performed a picture narrative task in English and rated the accentedness and comprehensibility of their own speech. These 41 speech samples were also rated for accentedness and comprehensibility by L1 English listeners, L1 Mandarin listeners, and listeners whose L1 was neither English nor Mandarin. Correlation and paired t-test analyses were conducted to investigate the relationships between self- and other-assessment. The key findings from Study 2 were: 1) L1 Mandarin speakers' self-assessment of their accentedness and comprehensibility was comparable to the judgement by listeners who are also L1 speakers of Mandarin; 2) L1 Mandarin speakers' self-assessment of their comprehensibility was comparable to the judgment by L1 mixed listeners, but the judgements of the accentedness of the Mandarin-accented speech by the two groups were significantly different, with the L1 Mandarin speakers perceiving their own speech to be less accented than the L1 mixed listeners did; 3) L1 Mandarin speakers' self-assessment of their pronunciation was significantly different from the judgement by L1 English listeners, with the speakers evaluating their own speech to be less accented and more comprehensible than the L1 English listeners did.

Together, Study 1 and Study 2 suggested that the alignment between L2 pronunciation self- and other-assessment may vary according to the L1 backgrounds of the speakers and listeners. To gain a better understanding of the patterns observed from

Study 1 and Study 2, Study 3 investigated if L2 listeners had an advantage over L1 listeners at comprehending L2 speech. Forty-one L1 Mandarin speakers performed a picture narrative task. The speech samples were later rated for comprehensibility by L1 English listeners, L1 Mandarin listeners, and L1 mixed listeners. Paired T-test analyses were conducted to determine if L2 listeners indeed had an advantage comprehending L2 speech. The key findings from Study 3 were: 1) speaking overall, L1 Mandarin listeners perceived the Mandarin-accented speech to be significantly more comprehensible than did the L1 English listeners; 2) when proficiency was taken into consideration, a comprehensibility benefit was again observed with LP-M speakers/LP-M listeners, LP-M speakers/HP-M listeners, and HP-M speakers/HP-M listeners; 3) overall speaking, L1 mixed listeners and L1 English listeners were comparable with each other in their perception of the comprehensibility of the Mandarin-accented English speech; and 4) when proficiency was taken into consideration, an interlanguage comprehensibility benefit was observed with LP-M speakers and LP-X listeners, whereas an interlanguage comprehensibility detriment was observed with HP-M speakers and LP-X listeners.

Conclusion and Implications

The findings suggest that L2 learners may not always have an accurate assessment of how accented or comprehensible their own speech sounds - a finding that is not surprising in light of the findings of Trofimovich et al. (2016) and a wealth of evidence from the field of social psychology. Additionally, the findings of this dissertation also indicate that L2 speakers' self-assessment accuracy may be related to both the L1

background of the listeners, and the L2 proficiency level of speakers themselves and the listeners. Helping L2 learners calibrate their self-assessment of L2 pronunciation should be a pedagogical objective for L2 educators. This may be achieved by reducing the ambiguity L2 learners experience with L2 pronunciation, which may include clarifying the constructs and expectations of pronunciation acquisition, providing easily recognizable negative feedback, and explicit information in terms of why failure has occurred. Additional methods suggested in Dunning et al. (2004) include review of past performance, benchmarking (comparing self-performance against that of others), and peer assessment.

Additionally, it is important for educators to think about the types of interactions the learners will engage in and prepare their students accordingly. For example, considering the tendency for high-proficiency speakers to under-evaluate themselves and the advantage L2 listeners may have comprehending L2 speech especially when the speakers and listeners share a common L1, it may be speculated that high-proficiency speakers are prone to under-evaluate themselves significantly when communicating with interlocutors with whom they share a common L1. With this type of learners, the pedagogical priority may be to help them avoid being preoccupied with language issues that are inconsequential to comprehensibility. Similarly, while low-proficiency speakers' self-assessment may align well with the judgement of L2 listeners from their own L1 background, or low-proficiency listeners from other L1s, their self-assessment may be overly inflated in other interactional environments. With this type of learners, educators

may need to help them recognize the different demands in different contexts, and set objectives accordingly.

Limitations and Future Directions

The findings of this dissertation brought to our attention once again that the L1 backgrounds of listeners and speakers affect how L2 comprehensibility and accentedness are perceived. Although the examination of the sources of the differences in self- vs. other-pronunciation assessment was not an objective of the current dissertation, a better understanding in this area will not only expand existing literature, but also provide educators useful information to adjust their instructional foci. Foote (2015) found that different speech variables were associated with comprehensibility judgement for interlocutors from different L1 backgrounds; and that for the L1 Mandarin listeners, an L1 effect above and beyond what could be explained by the properties of speech itself was found. It is recommended for future studies to examine the sources of the discrepancies observed in self- vs. other-assessment, and determine if/how the differences were derived from properties of the speech itself, or if there were other factors at play beyond what could be explained by speech characteristics.

While the current dissertation highlights the importance of calibrating L2 speakers' pronunciation self-assessment, it did not investigate how this may be practically achieved in L2 classrooms as well as the respective effectiveness of different approaches. Future studies in this direction are needed to provide educators the information and guidance needed to help learners in this respect.

Additionally, Study 1 indicated that speakers' L1 backgrounds may be a source of difference in terms of how well self-assessment may align with other-assessment. Study 2 and 3 suggested that the perception of L2 accentedness and comprehensibility may be closely related to the listeners' L1 background and their proficiency level. These findings highlight the importance of the inclusion of listeners from a variety of L1 backgrounds and at different proficiency levels in pronunciation research. These findings also suggest that research investigating L2 pronunciation assessment should avoid treating L2 speakers from different L1s as a homogenous group, or generalizing findings from studies that only utilize speakers or listeners from specific L1 backgrounds.

Additionally, as far as proficiency is concerned, existing studies investigating L2 listeners' assessment of L2 speech have used different criteria in the determination of high and low proficiency (e.g. accentedness, intelligibility, listening proficiency, self-reported general proficiency). Even among studies that adopted the same criteria (e.g. phonological proficiency), there may still be differences in the groups sampled as far as proficiency is concerned. This not only has made cross-study comparison difficult, but may also have been a source of the inconsistency in existing findings. It is recommended that future studies should bear this mind and approach the measuring of proficiency in a way that facilitates cross-study comparison.

APPENDIX

LANGUAGE BACKGROUND QUESTIONNAIRE

Age _____

Gender_____

Language Background

1. Place of birth (Country): _____
2. Your native language _____
3. Parents' native language(s) _____
4. Language(s) you use to speak with your family _____
5. Do you know any other language(s) besides your native language (e.g., beginner, intermediate, advanced, French, Spanish)?

   Language Proficiency

   - _____ _____
   - _____ _____
   - _____ _____

6. Have you ever had any hearing problems? (yes, no)_____
7. You are currently a:
   - Undergraduate student
   - Graduate student
   - Other. Please specify _____
8. Your major _____
9. How long have you been studying in the U.S.? _____years _____ months
10. Please indicate the countries in which you have lived for at least three months, the age you started living there, and how long you lived there.

    Country age length of stay

    _____ ____ _____

    _____ ____ _____

_____ ____ _____

11. How much time in total have you spent in the U.S. and other English-speaking countries (such as the U.K., Canada, Australia, etc.)? _____years _____months
12. At what age did you start learning English? _____
13. How many years have you been learning English? _____
14. In which context(s) have you learned English?
    - Primary school/elementary school
    - Middle school
    - High school
    - University
    - Language schools
    - Other. Please specify_____
15. Have you taken TOEFL iBT? (yes, no) _____

    Date of your TOEFL iBT test (month/year) _____
    - Your TOEFL iBT total score _____
    - Your TOEFL iBT listening score _____
    - Your TOEFL iBT speaking score _____
16. Have you taken IELTS? (yes, no) _____

    Date of your IELTS test (month/year) _____
    - Your IELTS total score _____
    - Your IELTS listening score _____
    - Your IELTS speaking score _____
17. What is your self-rating of your English speaking ability?

1	2	3	4	5	6	7	8	9
extremely low								extremely high

18. What is your self-rating of your English listening ability?

1	2	3	4	5	6	7	8	9
extremely low								extremely high

19. Approximately what percent of the time do you use English (speaking and listening) in your daily life? _____%
20. Of the time that you spend speaking English, approximately what percent of the time do you interact with native English speakers (as opposed to non-native speakers)? _____%
21. How familiar are you with Mandarin Chinese-accented English?

1	2	3	4	5	6	7	8	9
not at all familiar								very familiar

22. Other than Mandarin Chinese, are there any other foreign accents that are familiar to you? (yes/no)_____

    If "yes", please specify which foreign accent(s) and your level of familiarity (on a scale of 1-9, 1 = not at all familiar; 9 = very familiar)

    Language level of familiarity

    _____ _____

    _____ _____

    _____ _____
23. Are you taking any English classes this semester? If so, which class(es)?

_____

24. Are you going to take any English classes next semester? If so, which class(es)?

_____

Experience in Language Teaching and Training

25. Have you taught English? (yes, no) _____

    If "yes", where and how long have you taught English?

    _____ (location) _____ (duration)

    _____ (location) _____ (duration)

    _____ (location) _____ (duration)

26. Have you ever received any training on pronunciation teaching? (yes, no) _____
    If “yes”, what kind of training have you done? _____

Experience in Linguistics

27. Have you ever taken any linguistics classes (especially phonetics/phonology)? (yes, no) ____
    If “yes”, what kinds of classes?
    _____

BIBLIOGRAPHY

Alba, J. W., & Hutchinson, J. W. (2000). Knowledge calibration: What consumers know and what they think they know. *Journal of Consumer Research, 27*(2), 123-156.

Albrechtsen, D., Henriksen, B., & Faerch, C. (1980). Native speaker reactions to learners' spoken interlanguage. *Language Learning, 30*(2), 365-396.

Bachman, L. F., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing, 6*(1), 14-29.

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America, 114*(3), 1600-1610.

Berry, J. W., Phinney, J. S., Sam, D. L., & Vedder, P. (2006). Immigrant youth: Acculturation, identity, and adaptation. *Applied Psychology, 55*(3), 303-332.

Bongaerts, T., Van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition, 19*(4), 447-465.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707-729.

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America, 106*(4), 2074-2085.

Brantmeier, C. (2006). Advanced L2 learners and reading placement: Self-assessment, CBT, and subsequent performance. *System, 34*(1), 15-35.

Brantmeier, C., & Vanderplank, R. (2008). Descriptive and criterion-referenced self-assessment with L2 readers. *System, 36*(3), 456-477.

Breitkreutz, J., Derwing, T. M., & Rossiter, M. J. (2009). Pronunciation teaching practices in Canada. *TESL Canada Journal, 19*(1), 51-61.

Burda, A. N., Hageman, C. F., Scherz, J. A., & Edwards, H. T. (2003). Age and understanding speakers with Spanish or Taiwanese accents. *Perceptual and Motor Skills, 97*(1), 11-20.

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: how perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology, 90*(1), 60.

Calloway, D. R. (1980). Accent and the evaluation of ESL oral proficiency. In J. W. Oller, Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 102–115). Rowley, MA: Newbury House.

Carliner, G. (2000). The language ability of US immigrants: Assimilation and cohort effects. *International Migration Review*, 158-182.

Carter, T. J., & Dunning, D. (2008). Faulty Self-Assessment: Why Evaluating One's Own Competence Is an Intrinsically Difficult Task. *Social and Personality Psychology Compass, 2*(1), 346-360.

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching Pronunciation Hardback with Audio CDs (2): A Course Book and Reference Guide*. Cambridge University Press.

Chiswick, B. R., Lee, Y. L., & Miller, P. W. (2004). Immigrants' language skills: The Australian experience in a longitudinal survey. *International Migration Review, 38*(2), 611-654.

Chiswick, B. R., & Miller, P. W. (1995). The endogeneity between language and earnings: International analyses. *Journal of Labor Economics, 13*(2), 246-288.

Crystal, D. (2003). *English as a global language* (2nd ed.). Cambridge: Cambridge University Press.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America, 111*(6), 2862-2873.

Davidson, F., & Henning, G. (1985). A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. *Language Testing, 2*(2), 164-179.

Davila, A., Bohara, A., & Saenz, R. (1993). Accent penalties and the earnings of Mexican Americans. *Social Science Quarterly, 74*, 902-916.

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J., & Muhlenbruck, L. (1997). The accuracy-confidence correlation in the detection of deception. *Personality and Social Psychology Review, 1*(4), 346-357.

Derwing, T.M. (2003). What do ESL students say about their accents? *The Canadian Modern Language Review/La Revue Canadienne des Langues Vivantes 59*(4), 547–566.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379-397.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*(04), 476-490.

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation Fundamentals: Evidence-based perspectives for L2 teaching and research* (Vol. 42). John Benjamins Publishing Company.

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*(3), 393-410.

Derwing, T. M., & Rossiter, M. J. (2002). ESL learners' perceptions of their pronunciation needs and strategies. *System, 30*, 155-166. doi:10.1016/S0346-251X(02)00012-X

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*, 655-679.

Deterding, D., & Kirkpatrick, A. (2006). Emerging South-East Asian Englishes and intelligibility. *World Englishes, 25*(3-4), 391-409.

Dlaska, A., & Krekeler, C. (2008). Self-assessment of pronunciation. *System, 36*, 506-516.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69-106.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57*, 1082-1090.

Edele, A., Seuring, J., Kristen, C., & Stanat, P. (2015). Why bother with testing? The validity of immigrants' self-assessed language proficiency. *Social Science Research, 52*, 99-123.

Ensz, K. Y. (1982). French attitudes toward typical speech errors of American speakers of French. *The Modern Language Journal, 66*(2), 133-139.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395-430.

Fay, A. J., Jordan, A. H., & Ehrlinger, J. (2012). How Social Norms Promote Misleading Social Feedback and Inaccurate Self-Assessment. *Social and Personality Psychology Compass, 6*(2), 206-216.

Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in English sentences. *The Journal of the Acoustical Society of America, 84*(1), 70-79.

Flege, J., & Frieda, E. (1995). Amount of native-language (L1) use affects the pronunciation of an L2. *Journal of Phonetics, 25*, 169–186.

Flege, J. E., Munro, M. J., & MacKay, I. R. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America, 97*(5), 3125-3134.

Foote, J. (2010). *Second language learners' perceptions of their own recorded speech*. PMC Working Paper Series W P10-02 (pp. 3–27). Edmonton: Prairie Metropolis Centre.

Foote, J. A. (2015). *Pronunciation Pedagogy and Speech Perception: Three Studies* (Doctoral dissertation, Concordia University). https://spectrum.library.concordia.ca/980242/

Foote, J. A., Holtby, A. K., & Derwing, T. M. (2012). Survey of the teaching of pronunciation in adult ESL programs in Canada, 2010. *TESL Canada Journal, 29*(1), 1-22.

Foote, J. A., Trofimovich, P., Collins, L., & Soler-Urzúa, F. (2013). Pronunciation teaching practices in communicative second language classes. *The Language Learning Journal*. Advance online publication. doi: 10.1080/09571736.2013.784345

Fraser, H. (2010). Cognitive theory as a tool for teaching second language pronunciation. In S. De Knop, F. Boers, & A. De Rycker (Eds.), *Fostering language teaching efficiency through cognitive linguistics* (pp. 357-379). Berlin, Mouton de Gruyter.

Gardner, R. C., Tremblay, P. F., & Masgoret, A. (1997). Towards a full model of second language learning: An empirical investigation. *The Modern Language Journal, 81*(3), 344-362.

Gass, S. M., & Mackey, A. (2007). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 175–199). Mahwah, NJ: Lawrence Erlbaum.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34*(1), 65-87.

Glover, P. (2011). Using CEFR level descriptors to raise university students' awareness of their speaking skills. *Language Awareness, 20*, 121-133. doi:10.1080/09658416.2011.555556

Grant, L. (2014). *Pronunciation Myths: Applying second language research to classroom teaching*. Ann Arbor, MI: University of Michigan Press.

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing, 29*, 163-180.

Harmer, J. (2001). *The practice of English language teaching*. London/New York: Pearson Education ESL.

Hayes, A. F., & Dunning, D. (1997). Construal processes and trait ambiguity: Implications for self-peer agreement in personality judgment. *Journal of Personality and Social Psychology, 72*, 664-677.

Hayes-Harb, R., Smith, B. L., Bent, T., & Bradlow, A. R. (2008). The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts. *Journal of Phonetics, 36*(4), 664-679.

Heilenman, L. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing, 7*(2), 174-201.

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal, 70*(2), 125-132.

Imai, S., Walley, A. C., & Flege, J. E. (2005). Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by native English and Spanish listeners. *The Journal of the Acoustical Society of America, 117*(2), 896-907.

Ioup, G., Boustagui, E., El Tigi, M., & Moselle, M. (1994). Reexamining the critical period hypothesis: A case study of successful adult SLA in a naturalistic environment. *Studies in Second Language Acquisition, 16*(1), 73-98.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*(2), 135-159.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility. *Studies in Second Language Acquisition, 34*(3), 475-505.

Janssen-van Dieten, A. M. (1989). The development of a test of Dutch as a second language: The validity of self-assessment by inexperienced subjects. *Language Testing, 6*(1), 30-46.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics, 23*(1), 83-103.

Kachru, B. B. (1997). World Englishes and English-using communities. *Annual Review of Applied Linguistics, 17,* 66-87.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review, 64*(3), 459-489.

Krausert, S.R., 1991. *Determining the usefulness of self-assessment of foreign language skills: post-secondary ESL students' placement contribution*. Ph.D. Diss., University of Southern California.

Kruger, J., & Dunning, D. (1999). Unskilled or unaware of it: Difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121-1134.

Lambert, W. E., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *The Journal of Abnormal and Social Psychology, 60*(1), 44.

Lappin‐Fortin, K., & Rye, B. J. (2014). The Use of Pre-/Posttest and Self-Assessment Tools in a French Pronunciation Course. *Foreign Language Annals, 47*(2), 300-320.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly, 19*, 673-687.

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology, 46*(6), 1093-1096.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*, 369-377.

Levis, J. M. (2006). Pronunciation and the assessment of spoken language. In *Spoken English, TESOL and applied linguistics* (pp. 245-270). Palgrave Macmillan UK.

Little, D. (2008). Knowledge about language and learner autonomy. In J. Cenoz & N.H. Hornberger (Eds.), *Encyclopedia of language and education: knowledge about language* (Vol. 6, pp. 247–258). New York: Springer.

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). San Diego: Academic Press.

Lyster, R. (2001). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning, 51*, 265-301.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*(3), 280.

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language learning, 47*(2), 265-287.

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly, 36*(2), 173-190.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 075-100.

Millis, S. R., Jain, S. S., Eyles, M., Tulsky, D., Nadler, S. F., Foye, P. M., & DeLisa, J. A. (2002). Assessing Physicians' Interpersonal Skills: Do Patients and Physicians See Eye-to-Eye? *American Journal of Physical Medicine & Rehabilitation, 81*(12), 946-951.

Morley, J. (1991). The pronunciation component of teaching English to speakers of other languages. *TESOL Quarterly, 25*, 481-520.

Mouw, T., & Xie, Y. (1999). Bilingualism and the academic achievement of first-and second-generation Asian Americans: accommodation with or without assimilation? *American Sociological Review*, 232-252.

Moyer, A. (2013). *Foreign Accent: The Phenomenon of Non-native Speech*. Cambridge: Cambridge University Press.

Munro, M. J., & Derwing, T. M. (1994). Evaluations of foreign accent in extemporaneous and read material. *Language Testing, 11*(3), 253-266.

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

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49*, 285–310.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in second language acquisition, 28*(1), 111-131.

Nash, R. (1969). Intonational interference in the speech of Puerto Rican bilinguals, *Journal of English as a Second Language, 4*, 1–42.

Nelson, C. L. (2011). *Intelligibility in world Englishes*. Blackwell Publishing Ltd.

Noels, K. A., Clément, R., & Pelletier, L. G. (1999). Perceptions of teachers' communicative style and students' intrinsic and extrinsic motivation. *The Modern Language Journal, 83*(1), 23-34.

Noels, K. A., Pelletier, L. G., Clément, R., & Vallerand, R. J. (2000). Why are you learning a second language? Motivational orientations and self-determination theory. *Language Learning, 50*(1), 57-85.

Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing, 6*(1), 1–13.

Pallier, G. (2003). Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex Roles, 48*(5), 265-276.

Pendakur, K., & Pendakur, R. (2002). Language as both human capital and ethnicity. *International Migration Review, 36*(1), 147-177.

Pennington, M.C., & Richards, J.C. (1986). Pronunciation revisited. *TESOL Quarterly, 20*, 207-225.

Perlmutter, M. (1989). Intelligibility rating of L2 speech pre-and postintervention. *Perceptual and Motor Skills, 68*(2), 515-521.

Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29*(2), 191-215.

Politzer, R. L., (1976). Linguistic accuracy and intelligibility in *Proceedings of the 4th International Congress of Applied Linguistics* (Hochschul-Verlag, Stuttgart). pp. 505–513.

Raasch, A., 1980. Self-evaluation in adult education. (L'auto-evaluation dans l'enseignement des adultes). *Recherches et Echanges, 5*, 85–99.

Rogerson-Revell, P., & Miller, L. (1994). *Developing pronunciation skills through self-access learning. Directions in Self Access Language Learning*. Hong Kong University Press Hong Kong.

Ryan, E. B., & Carranza, M. A. (1975). Evaluative reactions of adolescents toward speakers of standard English and Mexican American accented English. Journal of *Personality and Social Psychology, 31*(5), 855.

Saito, K., & Shintani, N. (2016). Do native speakers of North American and Singapore English differentially perceive comprehensibility in second language speech? *TESOL Quarterly, 50*(2), 421-446.aito

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 3-32). Cambridge: Cambridge University Press.

Schmidt, R., & Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. *Talking to Learn: Conversation in Second Language Acquisition*, 237-326.

Scovel, T. (1988). *A Time to Speak: A Psycholinguistic Inquiry into the Critical Period for Human Speech*. Newbury House Publishers.

Scovel, T. (2000). A critical review of the critical period research. *Annual Review of Applied Linguistics, 20*, 213-223.

Selinker, L. (1972). Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching, 10*(1-4), 209-232.

Smith, L. E., & Bisazza, J. A. (1982). The comprehensibility of three varieties of English for college students in seven countries. *Language Learning, 32*(2), 259-269.

Smith, B., Bradlow, A. R., Bent, T., Solé, M. J., Recasens, D., & Romero, J. (2003). Production and perception of temporal contrasts in foreign-accented English. In *15th International Congress of Phonetic Sciences* (pp. 519-522). Causal Productions, Barcelona.

Smith, L. E., & Rafiqzad, K. (1979). English for cross-cultural communication: The question of intelligibility. *TESOL Quarterly*, 371-380.

Snyder, C. R., Higgins, R. L., & Stucky, R. J. (1983). *Excuses: Masquerades in search of grace* (No. 341). John Wiley & Sons.

Snyder, C. R., Shenkel, R. J., & Lowery, C. R. (1977). Acceptance of personality interpretations: the "Barnum Effect" and beyond. *Journal of Consulting and Clinical Psychology, 45*(1), 104.

Southwood, H., & Flege, J. (1999). The validity and reliability of scaling foreign accent. *Clinical Linguistics & Phonetics, 13,* 335-349.

Stibbard, R. M., & Lee, J. I. (2006). Evidence against the mismatched interlanguage intelligibility benefit hypothesis. *Journal of the Acoustical Society of America, 120*, 433–442.

Tauroza, S., & Luk, J. (1997). Accent and second language listening comprehension. *RELC Journal, 28*(1), 54-71.

The National Aeronautics and Space Administration, (2014). Pilot/Controller Communications. http://asrs.arc.nasa.gov/docs/rpsts/plt_ctlr.pdf from the NASA website 2/16/2016.

Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning, 41*, 177–204.

Thomson, R. I., & Derwing, T. M. (2014). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics, 36*(3), 326-344.

Tokumoto, M., & Shibata, M. (2011). Asian varieties of English: Attitudes towards pronunciation. *World Englishes, 30*, 392-408.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. Bilingualism: *Language and Cognition, 15*, 905-916.

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self-and other-perception of second language speech. *Bilingualism: Language and Cognition, 19*(1), 122-140.

Van Tubergen, F., & Kalmijn, M. (2009). A dynamic approach to the determinants of immigrants' language proficiency: The United States, 1980–2000. *International Migration Review, 43*(3), 519-543.

Van Wijngaarden, S. J. (2001). Intelligibility of native and non-native Dutch speech. *Speech Communication, 35*(1), 103-113.

Van Wijngaarden, S. J., Steeneken, H. J., & Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native listeners. *The Journal of the Acoustical Society of America, 111*(4), 1906-1916.

Varonis, E. M., & Gass, S. (1982). The comprehensibility of non-native speech. *Studies in Second Language Acquisition, 4*(02), 114-136.

Walker, R. (2010). *Teaching the pronunciation of English as a lingua franca*. Oxford: Oxford University Press.

Weinreich, U. (1953). *Languages in Contact: Findings and Problems*, NY: Linguistic Circle of NY, No. 1.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231-252.

Winters, S., & O'Brien, M. G. (2013). Perceived accentedness and intelligibility: The relative contributions of F0 and duration. *Speech Communication, 55*(3), 486-507.

Xie, X., & Fowler, C. A. (2013). Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English. *Journal of Phonetics, 41*(5), 369-378.

Yule, G., Damico, J., & Hoffman, P. (1987). Learners in transition: Evidence from the interaction of accuracy and self-monitoring skill in a listening task. *Language Learning, 37*, 511-521.

Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science, 9*(2), 111-125.

CURRICULUM VITAE