LINGUISTIC MEASURES OF SECOND LANGUAGE SPEECH: MOVING FROM MONOLOGIC TO INTERACTIVE SPEECH

By

Dustin Joseph Crowther

A DISSERTATION

Submitted to
Michigan State University
in partial fulfilment of the requirements
for the degree of

Second Language Studies – Doctor of Philosophy

2018

ABSTRACT

LINGUISTIC MEASURES OF SECOND LANGUAGE SPEECH: MOVING FROM MONOLOGIC TO INTERACTIVE SPEECH

By

Dustin Joseph Crowther

Second language (L2) scholars generally agree that pronunciation development should prioritize attaining understandable over nativelike speech (e.g., Derwing & Munro, 2015; Jenkins, 2000; Levis, 2005). What specific linguistic measures of speech enable listener understanding is less clear. While monologic-based research indicates a combined effect of segmental and suprasegmental measures (word stress, intonation, rhythm), interactive-based research has emphasized only a segmental focus. The current study takes a first step in addressing this divide by applying a monologic methodology to interactive speech.

Twenty intensive English program students (levels 3/4 of a 4-level program) completed one interactive and three monologic (Picture, Experiential, & Academic) tasks. Using 60-second (interactive) or 30-second (monologic) excerpts, 36 native Listeners rated each Speaker on 9-point scales per task for accentedness (i.e., nativelikeness) and comprehensibility (i.e., ease of understanding). I acoustically coded all utterances for a series of phonological and fluency measures (derived from Isaacs & Trofimovich, 2012). In addition, each Speaker received a task rating for the Experiential, Academic, and Interactive tasks to see if perceived accentedness and/or comprehensibility predicted actual task performance.

Consistent with previous findings, Listeners perceived comprehensibility more positively than accentedness (e.g., Trofimovich & Isaacs, 2012). In terms of task, Interactive speech patterned most similarly to Experiential speech, especially for comprehensibility. Speakers were easier to understand on these two tasks than they were for Picture or Academic. Across tasks,

Listeners' perception of comprehensibility was associated with fluency measures (Articulation Rate, Mean Length of Run), but not phonological. The more complex, linguistically constrained tasks, Picture and Academic, demonstrated stronger associations with these fluency measures than did Experiential and Interactive, a likely effect of the increased cognitive demands placed on Speakers in regards to their lexical retrieval and syntactical encoding processes (Segalowitz, 2010).

Listeners' perception of comprehensibility also associated with task performance on both the Experiential and Academic tasks, but not for the Interactive task. For both Experiential and Academic, it appears that a higher perceived comprehensibility rating aligns with higher overall task score (and for Experiential, scores in both the Pronunciation and Fluency categories). For Interactive speech, it is likely that task performance draws more upon measures of interactive competence (e.g., turn-taking, topic initiation, discourse extension; May, 2011) than it does perceived comprehensibility.

I conclude my study by discussing what insight the above findings can provide in regards to how L2 speech is perceived. This insight includes the potential effect of speaker, listener, and task variables, along with how the measurement of specific linguistic measures is operationalized. In addition, I discuss the potential pedagogical and assessment implications of perceived comprehensibility being associated with task performance. After addressing the limitations of my study, I provide suggestions for future research to extend my findings.

For mom, dad, and Tara, who have supported and encouraged me throughout all steps of my life journey thus far.

ACKNOWLEDGMENTS

My dissertation is the product of support from throughout my studies at Michigan State University. I begin by thanking everyone in the Second Language Studies program, including faculty, students, and alumni. Your support has gone beyond what can be expressed here alone.

Specifically, I would like to acknowledge the support my dissertation chair, Dr. Debra Hardison, has provided me throughout my four years of study. From time as an RA in my first semester to the last two years as a TA, having Debra's support has been a great benefit to my professional development. In addition, my dissertation committee members, Drs. Susan Gass, Peter De Costa, and Paula Winke, have provided guidance well beyond that necessary to complete my degree, and for that I am forever grateful.

I have been a long-time believer in the support that we as people provide each other. Thank you to my cohort: Jessica Fox, Susie Kim, Jie Liu, Jeff Maloney, Zack Miller, Magda Tigchelaar, and Irina Zaykovskaya. We have travelled these last four years together, and without your support along the way I would not be where I am today.

At different stages of my dissertation, many people have offered their time to help me along my way. Though not an exclusive list, I would like to acknowledge the help of Caitlin Cornell, Amanda Haag, Lizz Huntley, Dan Isbell, Minhye Kim, Jongbong, Lee, Jungmin Lim, and Susie Kim.

I would be remiss not to acknowledge the administrative support I have received throughout this process, and thus conclude by acknowledging MSU's English Language Center and Department of Linguistics and Germanic, Slavic, Asian and African Languages.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

ANOVA	Analysis of Variance
EIL	English as an International Language
ELF	English as a Lingua Franca
HCA	Hierarchical Cluster Analysis
ICC	Intraclass Correlation Coefficients
IELTS	International English Language Testing System
IEP	Intensive English Program
IRIS	A Digital Repository of Instruments and Materials for Research into Second Languages
L1	First Language
L2	Second Language
LFC	Lingua Franca Core
LRE	Language Related Episode
NNS	Nonnative Speaker
NS	Native Speaker
OPI	Oral Proficiency Interview
SLA	Second Language Acquisition
TESOL	Teaching English to Speakers of Other Languages
TOEFL	Test of English as a Foreign Language
TOEFL iBT	Test of English as a Foreign Language (internet-based test)
TOEIC	Test of English for International Communication

INTRODUCTION

A range of ideological perspectives have addressed second language (L2) pronunciation development, including Second Language Acquisition (SLA; Celce-Murcia, Brinton, & Goodwin, 2010), English as a lingua franca (ELF; Walker, 2010), and English as an international language (EIL; Low, 2015). A relatively consistent argument across these ideological views is that pronunciation instruction should adhere to what Levis (2005) referred to as the *Intelligibility Principle*, which places an emphasis on L2 learners' ability to be understood by their interlocutor(s). This is an extension beyond a longstanding focus solely on accent reduction (i.e., the *Nativeness Principle*). From an SLA perspective, this argument stems primarily from the fact that accented L2 speech is often unavoidable, even for L2 learners who begin at an early age (e.g., Abrahamsson & Hyltenstam, 2009; Flege, Munro, & MacKay, 1995; MacKay, Flege, & Imai, 2006; Major, 2001; Moyer, 2013). EIL and ELF scholars, who advocate on behalf of the ~75% of English users globally who are nonnative speakers (Crystal, 2008), prioritize a focus on achieving and maintaining mutual intelligibility over the attainment of any specific native-English norm (Seidlhofer, 2011). This is primarily due to the wide variety of L2 accents likely to be encountered during multilingual contact (Matsuda, 2017). While pronunciation instruction has been shown to be effective both at the phonemic (Lee, Jang, & Plonsky, 2015) and global (Derwing, Munro, & Wiebe, 1998; Saito & Saito, 2017) levels, the linguistic targets of such intervention-based instruction have varied, encompassing both segmental and suprasegmental measures of speech (Lee et al., 2015; Saito, 2012). To help complicate matters, respondents for surveys of actual classroom practice have indicated not only sporadic and unbalanced pronunciation instruction, but a preference for a segmental emphasis (e.g., Breitkreutz, Derwing, & Rossiter, 2001; Foote, Derwing, & Holtby, 2012; Hardison, 2014).

In line with a greater emphasis on intelligibility (i.e., understandable speech) over nativelikeness (i.e., accent-free speech) among speech production scholars (e.g., Derwing & Munro, 2015; Levis, 2005), I (along with my collaborators) have proposed that listener perception of L2 comprehensibility (i.e., ease of understanding) is associated with a wider range of linguistic dimensions (e.g., phonological, fluency, lexical, grammatical) than that of perceived accentedness (primarily phonological), with our findings based on both linguistic (coding of individual speech measures) and subjective (listener ratings of individual speech measures) assessments (e.g., Crowther, Trofimovich, Isaacs, & Saito, 2015a; Crowther, Trofimovich, Saito, & Isaacs, 2015b; Trofimovich & Isaacs, 2012). However, a primary focus on monologic tasks (e.g., picture narrative) has limited our ability to make any claims on L2 interactive performance. Considering that L2 usage in spontaneous communication often serves as an overarching goal of L2 acquisition (SLA; Loewen, 2015), and that it is within interaction that much SLA is theorized to occur (Gass & Mackey, 2015; Long, 1996), it is necessary to investigate whether the linguistic dimensions identified to promote understandable speech during monologic tasks are the same as those necessary for such performance during interaction.

Understanding within interactive speech is primarily considered through either researcher analysis of language-related episodes (LREs) and communicative breakdowns (e.g., Jenkins, 2000; Loewen & Isbell, 2017) or interlocutors' stimulated recall of communicative breakdowns (e.g., Gurzynski-Weiss & Baralt, 2014; Kennedy, Guénette, Murphy, & Allard, 2015). Such analyses have firstly emphasized lexical and grammatical issues over phonological, which supports our monologic argument that attaining mutual intelligibility requires more than just phonological accuracy. However, in instances where researchers/interlocutors have indicated phonological sources of communicative breakdowns, the emphasis has strongly been placed on

segmental rather than suprasegmental issues (e.g., Jenkins, 2000; Kennedy et al., 2015; Loewen & Isbell, 2017). This is in contrast to our monologic findings, which have argued for at least an equal, if not greater, role for suprasegmental measures in producing understandable speech. This difference in phonological emphasis between monologic and interactive understanding serves as the starting point for my dissertation.

A limitation of existing research on linguistic dimensions associated with understandable speech is a minimal focus on actual task performance. While we have drawn upon various tasks eliciting monologic speech, analyses have considered only listeners' perception of how accented or how understandable speakers are (usually through Likert scale ratings of ~30-second utterances), and not how effectively speakers have completed the actual task. As some tasks used to elicit speech feature readily available rubrics for task performance, such as those inspired by the International English Language Testing System (IELTS) or Test of English as a Foreign Language (TOEFL), this seems a serious limitation. Recent years have seen a greater emphasis on the relationship between L2 pronunciation and assessment (see Isaacs & Trofimovich, 2017, and Kang & Ginther, 2018, for two recent edited volumes), yet it is not clear how the pedagogically-orientated literature on L2 accentedness and comprehensibility that I draw upon (e.g., Derwing & Munro, 2015) may inform L2 assessment. Specifically, it is necessary to consider how phonology-based pedagogical targets drawn from the former actually inform the latter. As standards-based assessments (as contentious a topic as it is) play a significant role in L2 teaching and learning (e.g., Fulcher & Owen, 2016; Ginther & Elder, 2014), it would be a disservice to promote pedagogical targets that may promote understandable speech but do not necessarily lead to higher assessment performance.

CHAPTER 1: LITERATURE REVIEW

Throughout this chapter, I review relevant literature on the linguistic measures associated with listener perception of how accented and understandable second language (L2) speakers are, across both monologic and interactive speech. Within this review, I identify how methodological differences in the scholarship related to each speaking task type may explain diverging results. I then highlight potential divides between this more pedagogically-orientated body of L2 pronunciation research and how L2 speaking is viewed and addressed from an L2 assessment perspective. Finally, I present the six research questions that guide my dissertation.

Global Perception of Second Language Speech

Theoretically, L2 pronunciation has received relatively minor attention when it comes to models of L2 development and assessment (Galaczi, Post, Li, Barker, & Schmidt, 2017). While numerous variables attributed to L2 development in syntactic, lexical, and pragmatic dimensions (e.g., age, motivation, aptitude) have also been linked to pronunciation, rarely do empirical studies relate directly to theories of SLA (e.g., VanPatten & Williams, 2015). Instead, a primary focus of discussion has been on whether nativeness or intelligibility should be the primary target of L2 pronunciation acquisition (e.g., Derwing & Munro, 2015; Levis, 2005), and, recently, the linguistic dimensions (e.g., phonology, fluency, lexicon, grammar, discourse) associated with each (e.g., Isaacs & Trofimovich, 2012). Much research in this vein focuses on three key constructs, defined best in Derwing and Munro (2015):

- *Accentedness* – how distinguishable an L2 learners' speech pattern is from that of a member of the target speech community.
- *Comprehensibility* – how easy or difficult to understand a listener finds an L2 speaker's utterance to be.

- *Intelligibility* – how accurately a listener understands an L2 speaker's intended message.

A fourth construct, *fluency*, has also received a significant amount of attention, though much variation exists in how it has been defined (Chambers, 1997; Segalowitz, 2016). For example, Lennon (1990) makes reference to both 'broad' (global speaking ability) and 'narrow' (ease of delivery) conceptualizations of fluency. In line with the 'narrow' perspective, Derwing and Munro (2015) refer to the ease of flow of L2 speech, typically in reference to the presence/absence of pauses and other dysfluency markers. Segalowitz (2010) emphasizes the underlying processes involved with L2 fluency attainment, specifically addressing the link between cognitive (retrieval) and utterance (temporal) fluency. Importantly, Derwing and Munro (2015) have argued that these constructs, more specifically when comparing accentedness to either comprehensibility or intelligibility, are overlapping, yet partially independent. This is evidenced by the fact that L2 speakers can be perceived as both highly comprehensible/intelligible while still possessing a heavy accent (though a heavy accent is almost always present for speakers deemed to have low comprehensibility/intelligibility). While research within this stream has targeted primarily L2 English speech (e.g., Derwing & Munro, 2013; Isaacs & Trofimovich, 2012; Kang, 2010), recent years have seen an increased focus on additional languages, including Dutch (Caspers, 2010), French (Bergeron & Trofimovich, 2017), German (O'Brien, 2014), Japanese (Saito & Akiyama, 2017), Korean (Isbell, Park, & Lee, in press), and Spanish (Nagle, 2018).

The partial independence between global listener perception of accentedness and comprehensibility/intelligibility proposed by Derwing and Munro (2015) has been echoed in regards to the linguistic measures of L2 speech associated with listeners' perception of these same constructs (e.g., Crowther et al., 2015a; Crowther et al., 2015b; Isaacs & Trofimovich,

2012). Prior to Isaacs and Trofimovich (2012) and Trofimovich and Isaacs (2012), linguistic measures found to influence the perception of L2 speech had primarily been considered independent of each other, despite the range of measures that had been identified. For perceived accentedness, these associations included segmental accuracy (Derwing, et al., 1998), pausing and articulation rate (Trofimovich & Baker, 2006), and various suprasegmental measures such as pitch range, stress, and pause length (Kang, 2010). Associates of understanding (encompassing both comprehensibility and intelligibility) included word stress (Field, 2005) and speech rate (Munro & Derwing, 2001), as well as pitch range and pause or syllable length (Kang, Rubin, & Pickering, 2010; Winters & O'Brien, 2013). Despite having received less focus, non-phonological measures, such as grammatical accuracy, appear to play an important role in speech perception as well (Fayer & Krasinsky, 1987; Varonis & Gass, 1982). For example, Varonis and Gass (1982) asked native speakers (NSs) of English to rate the accent and comprehensibility of L2 speakers reading a pair of sentences, one of which was grammatical, one of which was not. Interestingly, they found that grammatical accuracy did indeed impact perceived accentedness, but only when the speaker was not seen as being at either end of the accentedness spectrum (i.e., highly accented vs. not-accented at all). Speakers were also perceived as being more comprehensible when reading grammatical, as opposed to non-grammatical sentences.

Drawing upon this knowledge, Isaacs and Trofimovich (2012) and Trofimovich and Isaacs (2012) employed correlational and regression analyses to measure the relative weight of strength of 19 linguistic measures on listener perception of L2 accentedness and comprehensibility. Participants were first language (L1) French/L2 English speakers completing a picture narrative, rated by 60 NSs of English using a pair of 9-point Likert scales. Results indicated that perceived accentedness was linked primarily to phonological measures, while

comprehensibility was associated with a wider range of considerations, now including fluency, grammatical, and lexical dimensions. A partial replication indicated similar results when the same speech data were rated by nonnative speakers (NNS) of English (L1 French, L1 Chinese; Crowther, Trofimovich & Isaacs, 2016). Further evidence for this distinction comes from Crowther et al. (2015b), where NSs rated the L2 English speech of learners from three distinct backgrounds (Chinese, Hindi, Farsi), as well as Crowther et al. (2015a), in which performance was compared across two tasks (IELTS- and TOEFL-inspired). Furthermore, recent evidence within this research agenda has strengthened the importance of lexicon in perceived comprehensibility (Saito, Webb, Trofimovich, & Isaacs, 2016). The message extending from this line of research, at least to this point, is that while perceived L2 accentedness is primarily associated with phonological measures, the rating of L2 comprehensibility requires listeners to draw on a much wider range of linguistic dimensions to attain understanding. In terms of phonological considerations, listener perception of comprehensibility has generally shown greater balance between segmental accuracy and suprasegmental measures (word stress, intonation, and rhythm) than for accentedness. Measures of fluency (articulation rate, mean length of run, pause accuracy) have generally been associated with both. An important consideration, however, is that such associations may be dependent on speakers' proficiency, as highlighted in Saito, Trofimovich, and Isaacs (2016). Considering the picture descriptions of three proficiency levels of L1 Japanese/L2 English speakers, the authors argued that for comprehensibility, fluency measures with varied prosody were most relevant for beginner-to-intermediate speakers. At the advanced level, segmental accuracy with good prosody was of greater importance. For accentedness, segmental and prosodic measures were important across levels, though grammaticality became of greater relevance at the advanced level. However, it

should be noted that Saito et al. operationalized proficiency through assigned comprehensibility ratings rather than any standardized measures of proficiency. This leaves an open question as to how reliable such findings may be in regards to such an association.

One important caveat to consider is an emphasis on comprehensibility (ease/difficulty of listener understanding), rather than intelligibility (accuracy of understanding). Though more closely related than either is with accentedness, comprehensibility and intelligibility are not 100% correlated (Derwing & Munro, 2015). A scholarly emphasis on comprehensibility has been due to the primary usage of scalar measures (e.g., 7- or 9-point Likert scales), which are more closely aligned with how "pronunciation" is often operationalized in high stakes assessments such as IELTS and TOEFL (Harding, 2017; Isaacs & Trofimovich, 2012). I place pronunciation above in quotations, as the constructs of accentedness and comprehensibility are quite often conflated within assessment rubrics (Harding, 2018; Isaacs, Trofimovich, Yu, & Muñoz Chereau, 2015). To maintain alignment with both the work of Isaacs and Trofimovich, as well as my own, I here maintain an emphasis on comprehensibility.[1]

From Monologues to Interaction

The linguistic-based partial independence between accentedness and comprehensibility described above is primarily derived from monologic speech, with a heavy emphasis on the usage of a picture narrative. Though still in need of further investigation, there is evidence that as task complexity increases and L2 speakers are forced to draw upon a wider range of their linguistic resources (Robinson, 2011; Skehan, 2009), the distinction between linguistic correlates of perceived accentedness and comprehensibility begins to diminish. Crowther, Trofimovich, Saito, and Isaacs (2017) compared the linguistic correlates of accentedness and

[1] However, as part of data collection procedures, I included a measure of intelligibility (via orthographic transcription) to be considered in future analyses.

comprehensibility across three tasks (Picture, IELTS- & TOEFL-inspired). Ten English NSs rated the accentedness and comprehensibility of 60 L2 English speakers, as well as speakers' performance on ten linguistic measures (spanning dimensions of phonology, fluency, lexicogrammar, discourse).[2] Drawing on both Robinson's Cognition Hypothesis (Robinson, 2005) as well as the speakers' own perceptions of task difficulty, Crowther et al. identified the TOEFL-inspired task (integrated speaking) as being more complex than the other two. Interestingly, while listeners' perceptions of accentedness associated only with phonological and fluency measures for the Picture and IELTS-inspired task, for the TOEFL-inspired task perceptions now also associated with grammatical and lexical measures. These associations aligned accentedness more closely with comprehensibility. Though more research is needed to explore the potential overlap between the two constructs in regards to task complexity, additional evidence exists for L2 French speech as well (Bergeron & Trofimovich, 2017).

If increasing task complexity across monologic tasks influences the perception of L2 accentedness and comprehensibility, it would seemingly follow that manipulating additional variables would further impact listener perception. Specifically, making a task + interactive opens up a number of additional variables to increase complexity. Drawing from Robinson's (2005) Cognition Hypothesis, the presence of an interlocutor could potentially impact task complexity in a number of ways. The Cognition Hypothesis includes a "Task Condition" category which features *participation* variables such as +/- open solution, +/- convergent solution, +/- few participants, and +/- negotiation not needed and *participant* variables such as +/- same proficiency, +/- same gender, +/- shared content knowledge, and +/- shared cultural knowledge. Just as monologic speaking tasks can differ in complexity, so too can interactive

[2] This study drew upon the same dataset used in Crowther et al., 2015a, 2015b.

speaking tasks. Yet, it is not known how the change from - interactive to + interactive may impact speakers' production and listeners' perception of linguistic dimensions of speech.

Aside from extending what has become a relatively influential stream of L2 pronunciation research, a focus on interaction may also help address the gap between existing L2 pronunciation research and theoretical views on SLA (Galaczi et al., 2017). Specifically, any consideration of the role of L2 accentedness and comprehensibility in interactive effectiveness would be remiss if it were to not consider the important role interaction plays in L2 development (Gass & Mackey, 2015).

Interaction Hypothesis. As put forth in the *Interaction Hypothesis* (e.g., Long, 1996), language learning occurs during communicative breakdowns in conversation involving L2 speakers. Within these breakdowns, in an effort to repair communication, speakers, whether native or non-native, will incorporate discourse moves such as clarification requests or comprehension and confirmation checks. These discourse moves comprise negotiation for meaning, which facilitates L2 development by drawing learners' attention to which linguistic measures of speech led to the communicative breakdown in question (see Gass & Mackey, 2015, and Mackey & Goo, 2007, for more in-depth breakdowns). Though discussing monologic tasks, Crowther et al. (2015a, 2015b) made a loose connection to the Interaction Hypothesis. As some linguistic dimensions are more likely to lead to communicative breakdowns than others (Mackey, Gass, & McDonough, 2000), we argued that identifying the linguistic measures of L2 speech related to perceived comprehensibility more so than accentedness would provide L2 learners with explicit knowledge to help them notice and repair their nontarget production during communicative breakdowns. Pedagogically, identifying phonological difficulties that need to be explicitly addressed would free up learners' cognitive processing to focus on the lexical and

syntactical measures that hinder communication. The issue, of course, is in the actual identification of these measures.

Linguistic sources of communicative breakdowns. Identifying the sources of communicative breakdowns in interaction has often been conducted through the analysis of language-related episodes (LREs), in which interlocutors discuss a linguistic item, either due to the breakdown itself or a desire/need for linguistic accuracy (Swain & Lapkin, 1998). However, pronunciation has rarely been the focus of such research (Loewen & Isbell, 2017). One recent example, however, comes from Kennedy et al. (2015), who considered the sources of communicative breakdowns between intermediate and advanced L2 French speakers in Québec, Canada. Using video recordings of each interaction (i.e., stimulated recall), Kennedy et al. asked interlocutors to comment on potential and actual comprehension problems. They found that 18% of reported comprehension issues were related to pronunciation, primarily due to segmental accuracy. Few links were made to suprasegmental measures of speech, and, interestingly, no effect of proficiency was found. These findings align with that of Loewen and Isbell (2017), who employed a more analytical approach, coding LREs for the linguistic measure of interest (minus the interlocutor input collected in Kennedy et al., 2015). Despite the different methodological approach, the results were strikingly similar, with 16% of LREs related to pronunciation (and 90% of these focused on segmental concerns). Though the emphasis on pronunciation-related LREs ranges from between 1% and 40%, (Bowles, Toth, & Adams, 2014; Bueno-Alastuey, 2013; Gurzynski-Weiss & Baralt, 2014), that greater than half of reported issues are attributed to non-phonology-based dimensions (e.g., lexicon, syntax) provides support to previous findings that listener understanding is as much reliant (in these examples more so) on grammatical and lexical considerations as it is phonological (e.g., Crowther et al., 2015a, 2015b; Isaacs &

Trofimovich, 2012). In terms of pronunciation, that segmental features appear to be the primary source of communicative breakdown aligns strongly with evidence from an ELF perspective.

Lingua Franca Core. To this date, one of the most well-cited analyses of potential pronunciation targets based on interactive data comes from Jenkins (2000, 2002). Jenkins proposed the Lingua Franca Core (LFC), a series of pedagogical targets designed to allow for mutual intelligibility between users of English from different linguistic and cultural backgrounds (with the caveat that despite the now common inclusion of NSs of English [Jenkins, 2014], ELF does not adhere to any specific native-English norm [Jenkins, 2006; Seidlhofer, 2011]). Through the observation of L2 speaker interactions (and "wherever feasible" [Jenkins, 2002, p. 87] discussion with the interlocutors) Jenkins identified pronunciation-based measures that hindered mutual intelligibility. Within the LFC, Jenkins advocates for core and non-core elements, which could be argued to align with Levis' (2005) Intelligibility (core) and Nativeness (non-core) Principles. A corpus-based approach, the LFC has been criticized due to its limited interlocutors and speech (Munro & Derwing, 2006; Sewell, 2017), and has been the source of much debate between ELF and non-ELF scholars (see Dziubalska-Kołaczyk & Przedlacka, 2005, for one example). Of potentially greater interest, however, is that the LFC emphasizes segmental accuracy (Park & Wee, 2015), in the process relegating suprasegmental measures (e.g., word stress, pitch range, and rhythm) to non-core status. While aligning with Kennedy et al. (2015) and Loewen and Isbell (2017) in regards to interactive tasks, this is in contrast to findings linking these suprasegmental features to comprehensibility on monologic tasks (Crowther et al., 2015a, 2015b; Dauer, 2005; Field, 2005; Isaacs & Trofimovich, 2012). It should be noted, though, that more contemporary ELF-inspired research has argued for a greater role of intonation in interactive meaning making (Pickering, 2009; Pickering & Litzenberg, 2011).

Methodological concerns. Clearly, methodological differences exist between monologic and interactive speech in terms of which phonologic measures hinder interlocutor understanding. Monologic-based data emphasizes listeners' impressionistic opinions across an entire utterance, whereas interactive-data relies on participants' retrospective analysis and outsiders' reflective observations of specific moments in time (i.e., communicative breakdowns). Similarly, the linguistic source(s) of these breakdowns are determined either through a fully-developed coding scheme that encompasses an entire sample (monologic) or the identification of the source of difficulty in specific episodes (interactive).

Within episodes of communicative breakdown, interactive analysis often emphasizes the perceptions of participants themselves, who, in turn, have placed an emphasis on segmental issues. Generally, segmental accuracy receives far greater emphasis in the L2 classroom (e.g., Foote et al., 2012; Foote, Trofimovich, Collins, & Urzúa, 2013; Hardison, 2014) which may bias L2 learners' metalinguistic awareness of phonological measures when engaging in LREs and stimulated recall. The existence of such a bias may help to explain the divergence between interactive and monologic findings. Similarly, monologic tasks target comprehensibility (i.e., perceived ease/difficulty of understanding). This may indicate the effort needed to understand an utterance, yet, does not inform us whether actual loss of meaning occurred. For interactive speech, the emphasis is on actual episodes of communicative breakdown, or moments where it is clear meaning was lost. As such, intelligibility (i.e., accuracy in which a speaker's intended message was understood) may be a more relevant construct for interactive tasks (Loewen & Isbell, 2017).[3] However, a primary focus on individual moments does not a) take into account forms less prevalent during target interactions (see Sewell, 2017), and b) allow for an

[3] As discussed earlier, it is important to remember that Derwing & Munro's (2015) constructs of comprehensibility and intelligibility are not perfectly aligned.

understanding of the internal processes learners engage in to understand speech throughout an interactive encounter (Oppenheimer, 2008); more specifically, how much effort they require to comprehend their speaking partner. This too may help to explain differences in phonological measures relevant to interactive versus monologic speech.

One way in which to begin to address these methodological gaps would be to apply a monologic approach to interactive tasks. As previously noted, comprehensibility, rather than intelligibility, was chosen for monologic studies as it was more aligned with the measure of understanding featured in high stakes assessment (Harding, 2017; Isaacs & Trofimovich, 2012). As high stakes assessments that include paired-tasks (e.g., Cambridge First Certificate of English) rely upon similar rubric-based measures (i.e., rater-based perceptions), gathering listener-based perceptions would be a logical first step of bridging the gap in methodology.

From Listener Perception to Task Performance

An additional concern prevalent in accentedness/comprehensibility-orientated research is that learners' actual task performance has often not been addressed, when in reality, this would seem to be the information most relevant to learners' SLA (i.e., do lower accentedness/ comprehensibility ratings impact learners' overall score on a monologic task? does a greater number of LREs lead to a lower score in an interactive task? are lower task scores related to specific linguistic measures?). If comprehensibility is chosen based on its alignment with high stakes assessments' measures of understanding (Harding, 2017; Isaacs & Trofimovich, 2012), then it seems reasonable that we should see some level of alignment between this global measure and task performance. However, if there is no alignment, then it may be that the linguistic measures identified from a listener-based perspective may be less relevant than those from a rater-based perspective.

Accentedness and comprehensibility in assessment rubrics. As has already been stated, accentedness and comprehensibility are partially overlapping constructs (Derwing & Munro, 2015). In fact, when correlations between the two have been reported, their strength of association can be quite high (for example, Crowther et al., 2016, found $r > .90$ for three different listening groups, and Crowther et al., 2017, found $r = .74$-$.80$ across three tasks).[4] While comprehensibility is generally rated higher than accentedness, an increase in one will still lead to an increase in the other. However, strength of correlation becomes a concern statistically as it may increase concerns of collinearity (Field, 2009), and ultimately, as the association between two constructs increases, there is concern on whether the two constructs are actually distinct or are simply a different measure of the same underlying skill (Warner, 2008), in this case L2 speaking. This may help to explain why, from an assessment perspective, accentedness and comprehensibility are often conflated into a single scale (e.g., Harding, 2018; Isaacs et al., 2015). For example, Ockey and French (2016) describe accent within their assessment framework as

> "the degree to which an individual's speech patterns are perceived to be different from the local variety, and how much this difference is perceived to impact comprehension of listeners who are familiar with the local variety. Therefore, *the strength of an accent indicates the degree to which it is judged to be different than the local variety, and how it is perceived to impact the comprehension of users of the local variety*" (p. 695, emphasis added).

In another example, Isaacs et al. (2015) describe how Derwing and Munro's (2015) constructs of accentedness, intelligibility, and comprehensibility have been conflated in the revised-IELTS pronunciation scale. For example, in Band 8, one descriptor states that pronunciation is "easy to

[4] Also see Munro & Derwing (1995), who included speaker-specific correlations and found that strength of association between accentedness and comprehensibility ranged from .41 to .82.

understand throughout; L1 accent has minimal effect on intelligibility." Isaacs et al. argue for greater precision in which construct is being addressed across Band scores, drawing upon the results of a mixed-methods analysis of eight IELTS examiners' use of the Pronunciation scale. Their results indicated inconsistent classification of examinees into Bands 5-8, with Bands 5 and 7 especially problematic. These bands also featured the least clear descriptors (e.g., "shows all the positive features of [band below] and some, but not all, of the positive features of [band below]." Focus group interviews additionally indicated that different examiners attended to different measures when conducting their ratings. Interestingly, despite variability between which measures examiners attended to, Isaacs et al. still found medium strength correlations between comprehensibility and IELTS Speaking (r = .51) and Pronunciation (r = .48) scores.

This tendency to conflate accentedness and comprehensibility into a single scale could be seen as somewhat prevalent in L2 pronunciation assessment (see also Harding, 2017, 2018, for recent overviews). Conflation of constructs is concerning, as the goals of Derwing and Munro's line of research (including my own) is pedagogically orientated, which should ideally serve L2 speakers well on such standardized tests as IELTS. However, if measures relevant to scale-based perception differ from those for rubric-based rating, there is clearly a potentially dangerous gap between L2 pronunciation instruction and assessment.

***A listener versus rater dichotomy*.** Before considering linguistic correlates of the rating scales used in tests such as IELTS, there is an important distinction to consider, described in depth by Yan and Ginther (2018). The evaluation of L2 speech production is usually conducted using two groups: listeners and/or raters. While *listener* groups can include nearly anyone, *rater* groups consist of those who have received formal training in how to rate speaking performance on a language proficiency test (e.g., IELTS, TOEFL). Primarily interested in impressionistic

judgments, research that employs *listeners* has emphasized global ratings of accentedness, comprehensibility, and intelligibility. From a listener perspective, measures of speech perception are often conducted through the use of Likert scale rating (accentedness, comprehensibility) or orthographic transcription (intelligibility; Derwing & Munro, 2015). In comparison, *raters*-based studies (e.g., Davis, 2009; Lazarton & Davis, 2008, Ockey, 2009) align with assessment goals of selection and placement (Yan & Ginther, 2018), and frequently employ some form of assessment rubric (e.g., IELTS, TOEFL) to collect perception measures.

Clearly, the body of research I have highlighted to this point relies heavily on listeners' impressionistic perceptions, as opposed to those of trained raters. While impressionistic perceptions of global measures are likely to be considered during rater scoring of L2 speech (Yan & Ginther, 2018), that listener-based studies have placed limited emphasis on overall task performance leaves the relative weight of impressionistic judgments on rater scoring unknown.

Linguistic correlates of rubric rating. Beyond the work of Isaacs et al. (2015) described above, little work has addressed the specific linguistic measures that raters attend to when providing task scores, whether for overall task performance or pronunciation-specific categories. Rather, research on monologic assessment has considered variables such as accent familiarity (e.g., Winke & Gass, 2013; Winke, Gass, & Myford, 2013), and paired assessment on discourse- and individual-based properties, such as interactive patterning (Galaczi, 2008), interlocutor personality (Ockey, 2009), and interactional competence (May, 2011). While recent studies have proposed rubrics geared towards assessing L2 learners' comprehensibility (Isaacs & Trofimovich, 2012; Isaacs, Trofimovich, & Foote, 2017), these rubrics draw upon listener-based rating data and focus group discussions with English for Academic Purposes professionals.[5]

[5] Both Isaacs & Trofimovich, 2012, and Isaacs et al., 2018, targeted L2 English for their scale.

They are not derived from any specific standardized assessment (e.g., IELTS, TOEFL). While Isaacs et al. (2017) propose their rubric as a tool for pre- and in-sessional university students, they still present it primarily as a tool to support continued oral language development.

Individual listener/rater effects. L2 pronunciation scholars, whether listener- or rater-orientated, have placed much attention on how individual characteristics may impact speech perception/rating, specifically in regards to linguistic training (e.g., Saito, Trofimovich, & Isaacs, 2017), familiarity with accented speech (e.g., Gass & Varonis, 1984; Winke & Gass, 2013; Winke et al., 2013), and the L1 background (native versus nonnative) of the listeners themselves (e.g., Bent & Bradlow, 2003; Harding, 2012; Major, Fitzmaurice, Bunta, & Balasubramanian, 2002).

Yan and Ginther (2018) distinguished listeners from raters based on formal training in a specific rubric. However, the notion of "training" is not foreign to listener-based research, though it manifests in a different way. Whether defined as "Trained vs. Naïve" or "Experienced vs. Inexperienced", listener-based research considers how linguistic training may impact speech perception, though such research has been inconclusive (e.g., Bongaerts, van Summeren, Planken, & Schils, 1997; Calloway, 1980; Saito et al., 2017; Thompson, 1991). Saito et al. (2017), working with the same data utilized in Isaacs and Trofimovich (2012) and Trofimovich and Isaacs (2012), found that a group of raters with linguistic and pedagogical experience provided more lenient ratings of accentedness and comprehensibility than their inexperienced peers, and were more consistent in evaluating complex linguistic measures (word stress, intonation, rhythm). This consistency led to Crowther et al. (2015a, 2015b, 2017) employing MA-level applied linguistics students with both L2 instructional and learning experience. They asked each listener to rate not only accentedness and comprehensibility, but also 10 linguistic

measures of speech (spanning phonology, fluency, lexicogrammar, and discourse dimensions). However, whether the perception of these experienced listeners aligned with those with less linguistic training was not addressed.

The effect of accent familiarity (whether L1 specific or nonnative in general) on task scoring has also been a popular empirical topic. Gass and Varonis (1984) found that for 142 English NS listeners, familiarity with a) nonnative speech in general, b) a specific nonnative accent, and c) a specific nonnative speaker facilitated their ability to draw accurate meaning from L2 utterances. Connecting back to the identification of linguistic associates, the "inexperienced" population (undergraduates) employed in Isaacs and Trofimovich (2012) and Trofimovich and Isaacs (2012) was located in the bilingual city of Montréal, Canada. So, while listeners may not have had formal linguistic training, they surely had significant exposure to both L1-French accented English and L1-English accented French. As with Crowther et al. (2015a, 2015b, 2017) above, how this accent familiarity may have impacted results is not clear. The role of accent familiarity has been further investigated in relation to a potential bias effect in L2 speech assessment (Winke & Gass, 2013; Winke et al., 2013), where such familiarity/bias has been identified as a potential source of compromise in test reliability between raters.

An accent familiarity advantage may not be limited to the NSs employed in the above studies, as evidence exists that NNS listeners/raters may have an easier time understanding same-L1 accented speech than their NNS peers who do not share the speaker's L1 (Bent & Bradlow, 2003; Harding, 2012; Major et al., 2002). However, contradictory findings (e.g., Crowther et al., 2016; Munro, Derwing, & Morton, 2006) have led to a belief that this advantage is only present for some NNS listeners/raters and only some of the time (Major et al., 2002; Munro et al., 2006), and that it may depend on variables such as L2 proficiency, context, and learner background

characteristics (Hayes-Harb, Smith, Bent, & Bradlow, 2008; Smith & Hayes-Harb, 2011; Xie & Fowler, 2013).

The interactive rubric. While oral proficiency interviews (OPIs) involve some degree of interactivity, they are still primarily learner-centered, as demonstrated by the existence of both person- and computer-moderated OPIs (e.g., Thompson, Cox, & Knapp, 2016). While interactive variety can indeed exist within such testing (e.g., Plough & Bogart, 2008), the role that the learner takes differs greatly than if they were to interact with a fellow learner. Learner-learner interaction elicits different discourse than when controlled by an examiner (e.g., Johnson & Tyler, 1998; Kormos, 1999). In fact, learners appear to perform better when engaging with a fellow learner than with a tester (Brooks, 2009). Paired assessment appears to receive far less focus in standardized testing, although the Cambridge First Certificate of English includes a 14-minute interactive component (http://www.cambridgeenglish.org/exams-and-tests/first/exam-format/; see Galaczi, 2008, for an in depth look into this assessment). Numerous variables have been considered in relation to paired assessment, all of which may impact the type and amount of language produced. These include pair/group dynamics (e.g., Galaczi, 2008; Storch, 2002), interlocutor proficiency (e.g., Csepes, 2009; Davies, 2009; Lazarton & Davis, 2009), interactional competence (May, 2011; Young, 2011), CAF measures (Sato, 2014), planning time (Niita & Nakatsuhara, 2014) and linguacultural differences (Scollon, Scollon, & Jones, 2012). However, the majority of rubric-based paired/group oral assessment scholarship has come from Ockey and colleagues (Ockey, 2009, 2011; Ockey, Koyama, Setoguchi, & Sun, 2015). In these studies, Ockey et al. have utilized a group-assessment rubric measuring fluency, grammar, vocabulary, pronunciation, and communicative strategies across five bands. While this stream of research has worked exclusively with university-level Japanese students, it has still proven

informative, specifically highlighting a high association between group oral performance and TOEFL iBT speaking scores (r = .76; Ockey et al., 2015). Of concern, and in line with Ockey and French (2014), is that the pronunciation category has conflated accentedness and comprehensibility into a single construct. For example, for Band 3 the following descriptor is provided: "Pronunciation is good but has still not mastered the sound system of English; accent does not interfere with comprehension." This once again makes it difficult to determine what specific dimensions raters are addressing, and if they are consistent in these dimensions.

The Current Study

My dissertation serves as a follow-up to my previous work (Crowther et al., 2015a, 2015b, 2016, 2017), which has aimed to identify pedagogical pronunciation targets that would prioritize L2 learners' ability to produce understandable speech (Derwing & Munro, 2015; Jenkins, 2000; Levis, 2005). However, the stream of research I have subscribed to, which focuses on the constructs of accentedness and comprehensibility, has prioritized monologic performance. As previously discussed, this line of research has produced findings that do not seem to align with those that have been found for interactive tasks. Specifically, and with a focus on listener understanding, monologic tasks place a greater emphasis on the production of suprasegmental measures, such as word stress, intonation, and rhythm (e.g., Crowther et al., 2015a, 2015b; Isaacs & Trofimovich, 2012). This is in contrast with interactive tasks, where listener understanding appears to be tied most significantly to segmental accuracy (e.g., Jenkins, 2000; Kennedy et al., 2015; Loewen & Isbell, 2017). This difference across task type may be related to two key methodologic differences. First, while monologic tasks emphasize comprehensibility (i.e., ease of understanding), intelligibility (i.e., accuracy of understanding an intended message) appears to be the primary focus of interactive analyses (Loewen & Isbell,

2017). Second, identifying linguistic correlates of comprehensibility on monologic tasks has been conducted through the coding of specific linguistic measures across longer utterances. In interactive tasks, researcher observation and interlocutor reflection of specific moments (i.e., communicative breakdowns) are used to identify sources of mis- or non-understanding.

To help bridge this gap, the current study applies monologic methodology to interactive speech. Twenty intensive English program (IEP) students completed one interactive and three monologic (Picture, Experiential, Academic) tasks. In the interactive task, speakers discussed an opinion-orientated topic with a fellow participant. Speakers participating came from one of two IEP levels, and represented two L1s, Japanese and Chinese, which allowed for some control over potential interlocutor effects.[6] Using 60-second (interactive) or 30-second (monologic) excerpts, NS listeners rated each speaker (on Likert scales) per task for accentedness and comprehensibility. I acoustically coded all utterances for a series of phonological and fluency measures (derived from Isaacs & Trofimovich, 2012). From this methodology, I address the following research questions:

1. Does listeners' perception of L2 accentedness and comprehensibility differ as a function of task (monologic vs. interactive)?
2. Do the linguistic measures of L2 speech that influence listeners' perception of L2 accentedness and comprehensibility differ as a function of task (monologic vs. interactive)?
3. Do listeners' perception of L2 accentedness and comprehensibility follow any patterns across task (monologic and interactive)?

[6] For transparency's sake, this was not an intentional choice, but the result of the population from which participants were drawn.

The next research question addresses one potential reason L2 learners may not reference suprasegmental measures during LRE- and stimulated recall-based analyses (if, of course, these measures are indeed tied to understanding). Teacher respondents to surveys on pronunciation instruction have indicated a segmental bias in the classroom (Breitkreutz et al., 2001; Foote et al., 2011, 2013; Hardison, 2014), which in turn may limit L2 learners' ability to articulate suprasegmental measures during LREs and stimulated recall. To gain a better understanding of L2 learners' knowledge of such measures, 29 IEP students completed a 5-point Likert scale questionnaire targeting familiarity with, previous instruction on, self-awareness of, and perceived importance of five phonological measures (consonants, vowels, word stress, intonation, rhythm). Drawing upon questionnaire results, I consider the following research question:

4. What awareness of phonological measures of L2 speech do learners possess?

Finally, as previously discussed, speech production has been measured from the perspective of both listener and rater (Yan & Ginther, 2018). While research questions 1-3 prioritize *listeners*, it remains to be seen how much impact such global ratings have on overall task performance, the primary target of *raters*. While it can be argued that accentedness and comprehensibility are all relevant to overall performance (Yan & Ginther, 2018), the relative weight of their impact is unknown. For this reason, three tasks (Experiential, Academic, Interaction) were assessed using task-specific rubrics. This allowed for the inclusion of research questions 5-6:

5. Does listeners' perception of L2 accentedness and comprehensibility predict overall task performance?
6. Do linguistic measures associated with listeners' perception of L2 accentedness and comprehensibility align with those associated with raters' task scores on monologic and interactive tasks?

My dissertation has theoretical, pedagogical, and assessment implications. Theoretically, L2 pronunciation has received relatively minor attention when it comes to models of L2 development and assessment (Galaczi et al., 2017). However, if we consider that L2 learning often occurs when learners' attention is drawn to linguistic measures of speech that lead to communicative breakdowns (i.e., the Interaction Approach; Long, 1996), of particular concern is whether the suprasegmental measures identified in monologic speech to create underlying listener difficulty are not perceivable during interaction. As interactive findings using LREs and stimulated recall indicate minimal interactive attention to suprasegmentals, it may be that such linguistic measures are in need of greater pedagogical focus. That segmental elements tend to receive greater classroom attention would support this argument (e.g., Breitkreutz et al., 2001; Foote et al., 2012, 2013; Hardison, 2014). Considering that explicit pronunciation instruction has been shown to be effective (Lee et al., 2015; Saito, 2012), addressing suprasegmental measures pedagogically would ideally minimize pronunciation as a concern during communicative breakdowns, further enabling L2 learners' existing preference to focus on lexical and grammatical targets. This proposal, though, is based on the hypothesis that the importance of suprasegmentals in producing understandable speech found during monologic performance is relevant to interactive performance.

In terms of assessment, accentedness and comprehensibility have often been confounded into a single scale (Harding, 2018; Isaacs & Trofimovich, 2012). While the generally high correlation between the two (e.g., Crowther et al., 2016, 2017) would serve as one justification for this, that comprehensibility is also usually found to significantly differ from accentedness (Derwing & Munro, 2015) indicates potential concerns with this approach. No study has yet to consider whether listeners' perception of L2 speakers' accentedness and comprehensibility,

conceptualized following Derwing and Munro (2015), inform task rating. If these constructs do indeed exert influence over task rating, then the pedagogical targets drawn from such research would serve to benefit both L2 pronunciation development and L2 assessment preparation. If not, then the targets identified in my previous studies (Crowther et al., 2015a, 2015b, 2016, 2017) may be limited in regards to their generalizable relevance.

CHAPTER 2: METHODOLOGY

Participants

Participants consisted of *Speakers* and *Assessors* recruited from the student body of an English-medium Midwest American university. The latter served as either *Listeners* or *Raters*, following Yan and Ginther (2018).

Speakers. I recruited 29 nonnative-English speakers (NNSs) from university-run intensive English program (IEP) courses. Students enrolled in IEP courses choose to pursue full-time English language study, with a school-designed placement test placing each into one of five proficiency levels (090-094). The 29 Speakers (M_{age} = 21.41 [SD = 4.87]; Female = 13, Male = 16) represented two L1 groups: Japanese (N = 15, female = 8, male = 7) and Chinese (N = 14, female = 5, male = 9).[7] Speakers began learning English on average at age 10.00 (SD = 3.08) and had studied for 11.46 years (SD = 4.50). Five Speakers reported prior study abroad experience in the US (1-3 years, all during high school). All but two Speakers reported English as their L2 (1 Japanese Speaker reported Chinese, 1 Chinese Speaker reported Japanese), and nine reported an L3 (Spanish = 3, Japanese = 2, Korean = 2, French = 2), albeit with only beginner's proficiency. An important difference is that while the Chinese Speakers enrolled in IEP courses with the goal of pursuing undergraduate study at the university, only one Japanese Speaker indicated a similar goal. The remaining 14 Japanese Speakers were participants on a semester-length study abroad, either company- (N = 5) or university-sponsored (N = 9). Table 1 provides biographical data, including standardized and self-assessed proficiency measures.

[7] Initial data collection included one L1 Vietnamese Speaker and one L1 Spanish Speaker. As all other Speakers were either Chinese or Japanese, I removed these two Speakers from analyses.

Initial data collection occurred during summer 2017 (N = 6). I recruited Speakers from the two highest levels of IEP, 093 (N = 4) and 094 (N = 2), via class visits. Speakers received US $20 as compensation, plus 60 minutes of English tutoring provided by me. The second round of data collection took place in fall 2017 (N = 23), with Speakers again recruited via class visits to IEP 093 (N = 1) and 094 (N = 22). Speakers received either US $20 plus 30-minutes of tutoring (N = 14) or 120-minutes of tutoring, but with no monetary compensation (N = 9)**.**

Assessors. Speakers performance was scored by either *Listeners* or *Raters*.

Listeners. Thirty-six native-English speaking undergraduate students (M_{age} = 20.61, SD = 1.20, Range = 18-25; Female = 35, Male = 1)[8] assigned speech scores for the L2 utterances of the 29 Speakers described above. I recruited Listeners from two of the university's TESOL minor courses (Second & Foreign Language Learning, Pedagogical Grammar), offered by the Department of Linguistics and Languages. Listeners were primarily education majors (N = 31), though additional majors included Spanish (N = 2), French (N = 1), Chinese (N = 1), and Linguistics (N = 1). Twenty-six reported pursuing the department's TESOL minor, with six also pursuing a language minor (Spanish = 5, Chinese = 1). Listeners had completed one of two *Intro to Linguistics* courses at the university but reported no additional theoretical linguistic courses. Those completing a Spanish major/minor had additionally taken several linguistic courses specific to their degree. Only three indicated prior language teaching experience, all with learners under 10 years of age (Spanish for 6 months in the US, English for one month in Japan, English for one month in Kazakhstan).

[8] The gender breakdown presented reflects that of the students enrolled in the department's TESOL minor program.

Twenty-One Listeners reported knowledge of an L2 (Spanish = 14, French = 2, Chinese = 2, Hindi = 1, Serbian = 1, American Sign Language = 1), and two of a third+ language (French & Spanish, Japanese).[9] Three had spent a short time abroad as part of their undergraduate studies (one semester in Ecuador [N = 2], one year in Spain). Listeners rated their exposure to specific accented-L2 English speech on 9-point Likert scales (1 = No previous exposure, 9 = Extensive previous exposure). Comparing the two primary L1s of the foci Speakers, Listeners reported greater familiarity with Chinese- (M = 4.06, SD = 2.40) than with Japanese- (M = 2.89, SD = 2.00) accented speech. A Wilcoxon signed-ranks test indicated this difference to be significant (Z = -3.26, p = .001), with a strong effect size (r = .61).

Speech rating occurred in late fall 2017. I recruited Listeners through class visits to TESOL minor courses with the permission of class instructors. As one course occurred online, the instructor forwarded a recruitment e-mail to enrolled students. All Listeners received class credit as assigned by their instructors. Due to the limited number of international students enrolled in the target TESOL minor courses (only four L1 Chinese students completed the procedure), no NNS Listener data are included. Though potentially interesting, scholars comparing NS and NNS listeners' ratings of the global speech measures to be targeted (i.e., accentedness, comprehensibility) have indicated minimal differences in perception between groups (Crowther et al., 2016; Derwing & Munro, 2013; MacKay et al., 2006).

Raters. Two NSs and two NNSs of English (M_{age} = 27.25, SD = 2.50; Female = 2, Male = 2) scored the task performance of the 29 Speakers. Raters were graduate students in second/foreign language teaching programs, and indicated relatively high levels of L2 familiarity

[9] Languages listed include only those in which participants rated their proficiency on a 9-point Likert scale as 2+. The scale end points were 1 = *Near beginner* and 9 = *near nativelike*.

both in general (8 [SD = .141, Range = 6–8]), as well as for Chinese- (7.50 [SD = 1.91, Range = 5–7]) and Japanese- (6.00 [SD = 2.94, Range = 3–9]) accented English speech, all measured on a 9-point Likert scale (1 = No previous exposure, 9 = Extensive previous exposure). All four Raters indicated previous language instructional experience, teaching English (N = 3), Arabic (N = 1), Chinese (N = 1), German (N = 1), and Latin (N = 1). On average, Raters had taught for 3.15 years (SD = 2.63), teaching a range of age groups (1.5-23) in both second (N = 2) and foreign (N = 4) language contexts. I recruited Raters from a graduate level L2 assessment course, and each received class credit as compensation. Further biographical data are provided in Table 2.

Materials

Monologic tasks. The three monologic tasks were the same as those used in Crowther et al. (2015a, 2015b, 2017). They consisted of a picture narrative (hereafter referred to as *Picture*), an IELTS-inspired long turn task (hereafter *Experiential*), and a TOEFL iBT-inspired integrative task (hereafter *Academic*). Having been used in previous research, the three tasks were established speech elicitation tools, and allowed for comparison across studies. It is important to note that while IELTS and TOEFL iBT inspired the Experiential and Academic tasks respectively, the same stringent procedures utilized in high stakes assessment were not present during data collection. For this reason, I have chosen to use more descriptive labels throughout.

Picture. The Picture task was the same used in much previous speech production research (e.g., Derwing, Munro, Thomson, & Rossiter, 2009; Isaacs & Trofimovich, 2012), and is available through the IRIS Database (Marsden, Mackey, & Plonsky, 2016) under Derwing et al. (2009). The eight-framed colored picture narrative depicts a story of two strangers who bump into each other on a busy street corner, and in the process accidentally exchange their identical suitcases. Upon returning home and opening their suitcase, they realize their mistake. Following

Table 1

Biographical data for Speakers.

N	Age			Age of Onset	Years of Study	Proficiency (SD)		Self-Rated Proficiency (SD) (1 = low ability, 9 = high ability)			
	Mean (SD)	Median	Range	(SD)	(SD)	TOEFL (N = 18)	TOEIC (N = 12)	Speaking	Listening	Reading	Writing
29	21.41 (4.87)	20	18-36	10.00 (3.08)	11.46 (4.50)	71.44 (5.37)	678.75 (90.73)	4.72 (1.49)	5.27 (1.38)	5.66 (1.22)	5.23 (1.23)

Notes. N = Sample Size; SD = standard deviation.

Table 2

Biographical data for each Rater.

Rater #	Age	L1	L2	L2 Proficiency	Current Degree (Field)	Teaching Experience	Accent Familiarity	
							Chinese	Japanese
1	27	German	English	TOEFL (101)	MA (German Studies)	0.5 years (Latin) 0.5 years (German)	7	4
2	28	Arabic	English	TOEFL (102)	MA (TESOL)	4 years (Arabic) .75 years (English)	5	3
3	30	English	Chinese	OPIc (Advance Mid)	MA (TESOL)	5.5 years (Chinese) 0.5 years (English)	9	8
4	24	English	Japanese	OPIc (Intermediate High)	MA (TESOL)	10 weeks (English)	9	9

Notes. 1 – accent familiarity rated on a 9-point scale (1 = Not familiar at all, 9 = Very familiar).

standard procedures (e.g., Derwing et al., 2009; Isaacs & Trofimovich, 2012), I provided Speakers one minute to preview the eight pictures before they provided their response. The picture narrative can be found in Appendix A.

Experiential. The Experiential task drew upon two publicly available IELTS prompts which required Speakers to discuss a prior experience in their life. The first version asked participants to *describe a party that they enjoyed* (International English Language Testing System, 2009), the second to *describe a restaurant that they enjoyed going to* (International English Language Testing System, 2011). Each Speaker received a card with their written prompt, along with several suggestions of discussion points. They had up to 1 minute to prepare their response (notes were allowed) before they spoke for between 1–2 minutes. Acting as the moderator, I followed up each response with one or two questions (e.g., *Have you been to any other similar parties?* for the prompt about describing a party). Appendix B provides the full prompts for both versions.

Academic. The Academic task made use of two TOEFL prompts, publicly available through sample test materials (Educational Testing Service, 2012), and targeted skills deemed necessary for successful academic study. For each prompt, Speakers had 45-50 seconds to read a short passage, before listening to an audio recording on a related topic. Upon completion of the audio recording, Speakers responded to a question related to the content of the two sources of input. They had 30 seconds to prepare a response (notes were allowed) and then spoke for one minute, drawing on examples from both the reading and audio when formulating their response. An audio-recorded examiner moderated the task, presented via a PowerPoint presentation.[10] The topic of Version A was *social interaction* (104-word text, 95-second audio) and the topic of

[10] The same sample test materials provided the sound file for the audio-recorded examiner, which I embedded within the PowerPoint.

Version B was *cognitive dissonance* (88 words, 80 seconds). Written and audio text are available in Appendix C.

Interactive task. The work of Ockey and colleagues (2009, 2011, 2015) motivated the Interactive task materials. In these studies, the authors assessed L1 Japanese university-level, English as a foreign language learners through a group oral discussion. Topics across the studies varied but were generally open in regards to how learners might respond. As all my participants were drawn from the university's IEP program, I chose to pursue academic-themed topics.

Pilot interactions with IEP students in spring 2017 indicated that my initial academic-specific prompts were not appropriate across cultural backgrounds. For example, the pilot prompt "It is better to select a university major prior to beginning OR after completing a year of study" proved troublesome for Speakers from cultures that were not afforded such choices. As such, I selected more extra-curricular-based topics for use. The final discussion prompts requested Speakers to agree or disagree with one of three statements: it is important to a) *attend many activities when studying abroad*, b) *make international friends when studying abroad*, and c) *travel to many places when studying abroad*. Prompts were counterbalanced across dyads.

A series of prompt questions accompanied each statement (e.g., a] have you attended any new activities while at the school? b] do you want to make international friends while here at the school? c] have you visited anywhere while at the school?). Directions informed Speakers to first express their opinion to their partner and then determine if their opinion differed from their partner's (and persuade their partner of their opinion if a difference existed). Speakers had 2 minutes to prepare for the interaction but were not allowed to take notes. The full interactive prompts are available in Appendix D.

Task comparisons. Crowther et al. (2017) differentiated the three monologic tasks using Robinson's (2005) Cognition Hypothesis. Their categorization is reproduced in Table 3. The authors deemed Academic (referred to as TOEFL) as more complex than Picture and Experiential (referred to as IELTS) due to the greater reasoning demands of the task. Participant ratings indicated a similar perception, with Academic seen as more complex than both Picture (p = .009) and Experiential (p = .005). Beyond Robinson's Cognition Hypothesis, the authors differentiated between Picture and Experiential based on the greater linguistic constraints placed on participants when presenting a picture narrative. Participants are constrained by the lexical items required to complete the narrative, whereas in the Experiential task they are free to draw upon their entire linguistic repertoire.

Table 3

Task complexity across three monologic tasks (as reported in Crowther et al., 2017).

	Picture	Experiential	Academic
Few elements	+	+	–
Spatial reasoning	+	+	–
Here/now	+	–	–
Casual reasoning	–	–	+
Intentional reasoning	–	–	+
Perspective taking	–	–	+

Notes. Complexity categories drawn from Robinson, 2005.

The Interactive task is potentially more complex due to the presence of an interlocutor (+ Interactive). However, as applied in the current study, the overall complexity of the interactive task may be limited, as, following Robinson's (2005) Cognition Hypothesis, the task is + open solution, + convergent solution, and + few participants. The complexity would come from – one-way flow, – few contributions needed, and – negotiation not needed. Ultimately, how complex

the task was depended on whether Speakers aligned or differed in their response and how often they felt compelled to contribute. As this was not an assessment context (as it was in the Ockey studies), there were no consequences if a Speaker chose not to fully engage. It should be noted that Robinson's participant variables (e.g., +/– same proficiency, +/– same gender) were not considered in the current study, due primarily to the relative homogeneity of the Speakers recruited, and their subsequent dyads (reported later). In terms of linguistic resources, as prompts were opinion-based, Speakers had the freedom to draw upon linguistic resources they felt would best support their intended message. This is similar to the linguistic freedom available in the Experiential task. The primary difference between the two is that, ideally, Speakers would take into consideration the contributions of their partner. Figure 1 presents a continuum of linguistic constraint across the four tasks.

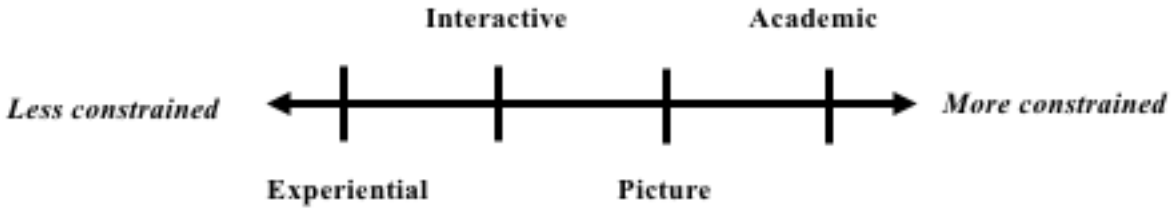

Figure 1. *Continuum of linguistic constraint across 4 speaking tasks.*

Pronunciation survey. Speakers completed a pronunciation survey which targeted six key phonological/fluency measures of L2 speech. Based on the rating categories of Crowther et al. (2015a, 2015b), the measures included Segments (divided into Consonants and Vowels), Word Stress, Intonation, Rhythm, and Speech Rate. Each measure received a brief written explanation, before Speakers used a series of 5-point Likert scales to rate their familiarity with each, the amount of instruction they had received, their self-awareness of the measure when speaking, and their perceived importance of the measure for intelligible speech. Five NNSs piloted the survey, and an experienced TESOL practitioner provided additional feedback. Based

on the feedback provided, I clarified each of the written explanations. Appendix E provides the complete survey, including end-point descriptors for the 5-point scales.

Background questionnaire. Speakers filled-in Questionnaire A (Appendix F), which targeted biographical information, language learning history, study abroad history, and university study plan. Listeners and Raters completed Questionnaires B and C (Appendices G and H) respectively, both of which requested biographical information, education history, language learning and teaching history, and accent familiarity.

Procedure – Speech Elicitation

Speakers committed a total of 60 minutes to data collection: 30 minutes for monologic task completion and 30 minutes for interactive task completion. Data collection occurred in 90-minute blocks involving two Speakers each. *Speaker 1* would arrive first and complete the 30-minute monologic session. As this session finished, *Speaker 2* would arrive and both would engage in the 30-minute interactive session. Upon completion, *Speaker 1* would leave and *Speaker 2* would complete their own 30-minute monologic session. This approach also enabled a counterbalancing of monologic and interactive task performance. *Speaker 1* read and signed a consent form before beginning their monologic session, *Speaker 2* before their interactive session.

Monologic session. The Speaker and I completed Questionnaire A together, allowing me to ask follow-up and clarification questions when needed. Speakers then completed the three monologic tasks, with me serving as their moderator to provide instruction and clarification when needed. The order of tasks was counterbalanced (e.g., Picture–Experiential–Academic; Academic–Picture–Experiential, etc.) to control for any potential task ordering effect. A Sony

ICD-PX333 digital voice recorder recorded all Speaker output. Upon completion, the Speaker and I completed the pronunciation survey, with clarification provided when requested/necessary.

Interactive session. The two Speakers met in a large room with two chairs positioned 2-3 feet apart. After introductions, I explained that they would each receive the same interactive prompt, which would serve as the basis for an 8- to 10-minute audio and visual recorded interaction. Each Speaker then took two minutes to read through the prompt and prepare a response. Before beginning, I addressed any clarification questions, and switched on the audio (Sony ICD-PX333 digital recorder) and video (Sony HDR-CX580 camera) recorders. Speakers then interacted. The completion of the interaction occurred either organically as the two Speakers appeared to have nothing left to discuss or deliberately by me after 8-10 minutes, at an appropriate place in the interaction (e.g., completed thought, short pause). Speakers then completed the post-interaction questionnaire. Interaction length ranged from 4:02-8:31 (*M* = 6:36, SD = 1.31).

Procedure – Speech Rating

Stimuli preparation. I prepared each monologic and interactive speech sample for speech rating.

Monologic. In line with Crowther et al. (2015a, 2015b), I normalized speech samples for peak amplitude and edited each down to the initial 30-seconds of speech produced, removing all initial disfluencies (e.g., uh, um) and false starts. This length falls in line with previous speech production research using 20- to 60-second recordings to elicit listener judgments of L2 speech (Derwing, Munro, & Thomson, 2008), while also being long enough to allow for reliable judgments (Munro, Derwing, & Burgess, 2010). In addition to these 30-second excerpts, I identified two approximately 10-second excerpts, with logical beginning and end points, per

monologic task. Intended as measures of intelligibility (i.e., accuracy of understanding) in a future study, I will not discuss these utterances in the following analyses. I refer to them here only to make clear the procedure Listeners went through, described below. In summary, for each monologic task, each Speaker provided one 30-second and two approximately 10-second excerpts.

Interactive. I reviewed each interactive speech sample to identify a 60-second excerpt that prominently featured both Speakers. Across interactive samples, speaking time ranged from 20.52s–38.4s (Mean = 29.60, SD = 5.03, Median = 31.30). Though not balanced, no excerpt involved a Speaker speaking for less than 37% of the time. This provided Listeners with at least 20 seconds of speech, allowing for reliable judgements (Derwing et al., 2008; Munro et al., 2010). Again, I normalized each sample for peak amplitude and removed all initial disfluencies.

Speech rating. Listeners completed speech rating through two 60-minute online questionnaires using Qualtrics (www.qualtrics.com). In Questionnaire 1, Listeners used 9-point Likert scales to rate each monologic speech utterance for accentedness and comprehensibility. In Questionnaire 2, Listeners orthographically transcribed a subset of the 10-second utterances as a measure of intelligibility and rated each interactive sample for accentedness and comprehensibility. Though the majority of global speech rating studies tend to collect speech ratings on site (e.g., Crowther et al., 2015a, 2015b; Munro & Derwing, 1999), collecting ratings through online measures have still provided high inter-rater reliability (Crowther et al., 2016). Three NSs drawn from the same target population piloted monologic and interactive rating procedures for timing, clarity of direction, and appropriateness.

Monologic. Listeners rated each 30-second recording for accentedness and comprehensibility using 9-point Likert scales. Despite some debate on appropriateness (Isaacs &

Thomson, 2013; Isbell, 2018; Munro, 2018; Southwood & Flege, 1999), the use of 9-point scales aligned with much previous research in the area (e.g., Derwing et al., 2015; Isaacs & Trofimovich, 2012). Following Derwing and Munro (2015), I informed Listeners to treat accentedness as the degree of difference between the Speaker's variety of English compared to the target variety (1 = *heavily accented,* 9 = *not accented at all*), and comprehensibility as how much effort they required to understand the utterance (1 = *hard to understand*, 9 = *easy to understand*). Listeners heard each utterance once. They could not advance to the next item until they had both heard the entire 30-second recording and provided a rating for both accentedness and comprehensibility (though these ratings could be provided at any time during the recording). While the choice to perform both ratings at once is motivated by time considerations, O'Brien (2016) found minimal differences between rating both constructs at once or individually.[11] Figure 2 presents the Qualtrics interface.

Listeners received online instruction before beginning speech rating. This included a written explanation of each of the targeted constructs and three practice ratings (using pilot recordings). For accentedness and comprehensibility, the Qualtrics interface informed Listeners that each recording would end after 30 seconds, potentially cutting a Speaker off mid-sentence and that this should not be considered in their rating. Following Crowther et al. (2015a, 2015b), Listeners rated the three tasks in counterbalanced blocks (e.g., Picture–Experiential–Academic; Experiential–Academic–Picture, etc.), with recordings within each block randomized into one of six possible orders.

[11] O'Brien (2016) included a third construct, fluency. She did find that L1 German listeners rating L2 German speech indicate greater fluency when rating all three constructs at once than when rating them individually (p = .018). There was also a trend indicating that L1 English-L2 German listeners rated L2 German speech as slightly more comprehensible when rating all three constructs together (p = .054).

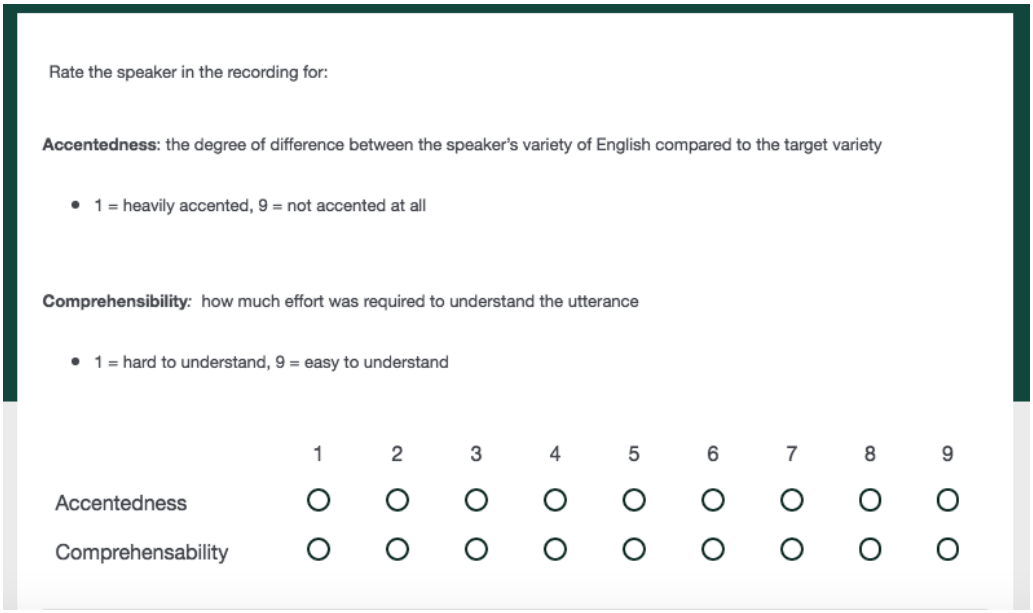

Figure 2. Qualtrics interface for monologic task rating.

Interactive. Listeners rated both Speakers within the 60-second interactive samples simultaneously, using the same 9-point scales for accentedness and comprehensibility. The Qualtrics interface informed Listeners that the first voice they heard would be considered *Speaker A* and the second voice *Speaker B.* As shown in Figure 3, the 9-point scales designated for Speaker A and Speaker B were clearly labelled. In the case of Speaker confusion, the interface provided Listeners the option to indicate that they were unable to clearly differentiate between the two Speakers. After completing their accentedness and comprehensibility ratings for both Speakers, Listeners moved to the next sample. Listeners completed Interactive speech rating in one of seven randomized blocks. As with the monologic tasks, Listeners received online instruction prior to rating, including three practice ratings.

At the end of the two rating sessions, Listeners self-rated their understanding of (1 = *I did not understand at all*, 9 = *I understand this concept well*) and comfort with (1 = *very difficult*, 9 = very *easy and comfortable*) rating both accentedness and comprehensibility (Appendix I).

Listeners indicated greater understanding ($Mean_{Acc}$ = 7.58 [SD = 1.71], $Mean_{Comp}$ = 7.81 [SD = 1.60]) than they did comfort ($Mean_{Acc}$ = 6.19 [SD = 1.78], $Mean_{Comp}$ = 6.89 [SD = 1.67]).

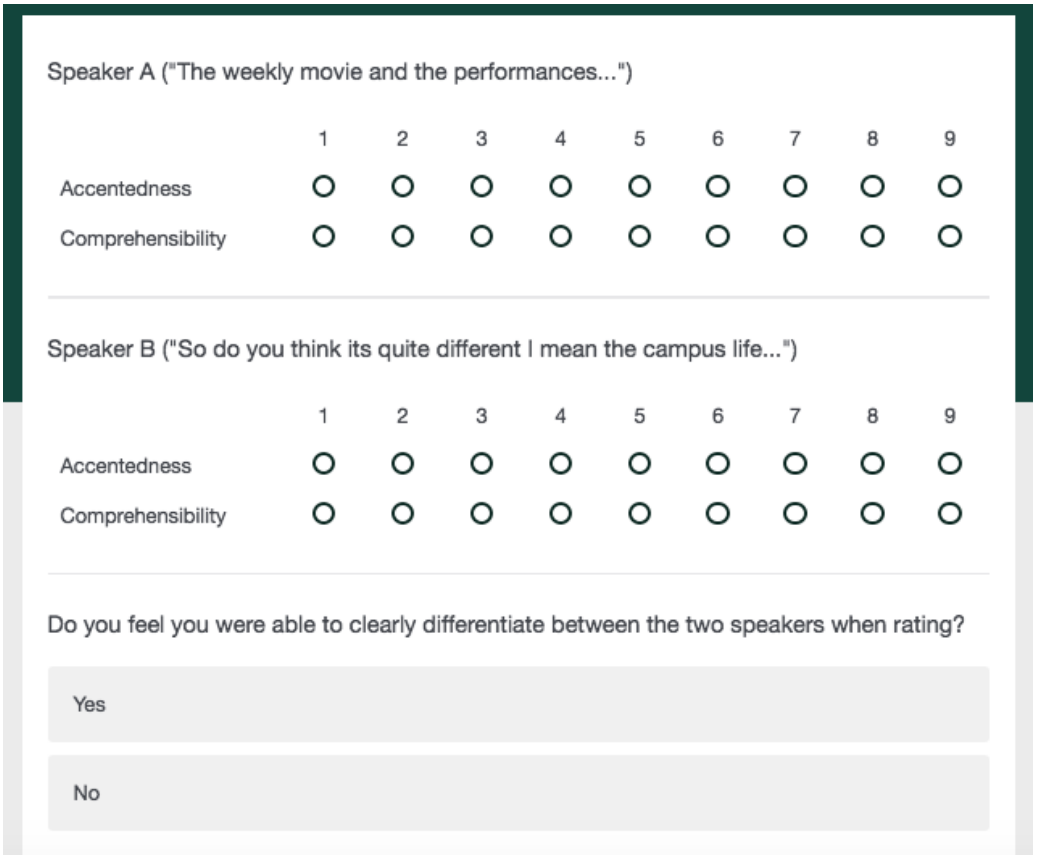

Figure 3. Qualtrics interface for interactive task rating.

Task scoring. The four Raters assessed Speakers on their Experiential, Academic, and Interactive performance. Rating occurred using the audio recordings of monologic performance and video/audio recordings of interactive performance. Two Raters rated Experiential while the other two rated Academic. The four Raters worked in their monologic pairs to each rate half the interactive task samples.

Each pair worked collaboratively to reach consensus on a task score for their assigned Speakers. For training on monologic rating, each pair met with me for a 45-minute session. Prior to training, each Rater took a week to familiarize themselves with their rubric. The training

session consisted of three stages. In Stage 1, each pair expressed and discussed questions and concerns regarding their assigned rubric. In Stage 2, they worked collaboratively to rate two speech samples determined by me to represent a high performing and low performing Speaker. The final stage had each Rater work individually to rate 2-3 additional samples, before comparing ratings with their partner. Raters then discussed and resolved any discrepancies on their own. Interactive task training followed the same procedure, except that all four Raters were present. Upon completing training, Raters took a month to collaboratively assign a score per Speaker per task.

Monologic. Experiential and Academic task scoring used predesigned rubrics. The Experiential task used the publically available IELTS speaking rubric (International English Language Testing System, 2016). This 10-band rubric (0-9) featured four categories (fluency and coherence, lexical resource, grammatical range and accuracy, pronunciation). Raters determined an overall score by tallying band scores across categories and dividing by four (the number of categories). Scores of .25 or .75 were rounded up to .5 and .00 respectively (e.g., 4.25 -> 5.5, 4.75 -> 5.0). For example, Speaker #3 received scores of 5 (fluency and coherence), 6 (lexical resource), 6 (grammatical range and accuracy), and 6 (pronunciation). Dividing their summed score (23) by number of categories (4), their overall Experiential score was 5.75, which was rounded up to 6. Although this procedure emulated that utilized for official IELTS rating (https://www.ielts.org/en-us/ielts-for-organisations/ielts-scoring-in-detail), it did not feature the same level of training rigor.

The Academic task employed the publically available TOEFL integrated speaking rubric (Educational Testing Service, 2014). This 5-band rubric (0-5) featured four categories (general description, delivery, language use, topic development). Raters provided a single band score

representative across categories. As with the Experiential task, scoring followed general TOEFL procedures (https://www.ets.org/toefl/ibt/scores/understand/), though Raters did not possess the same rigor of training. For both rubrics, descriptors accompanied each band for each category.

Unlike Experiential and Academic, no rating rubric exists for this specific Picture task. A search of picture narrative tasks returned limited assessment options (e.g., Sato, 2014, which targeted only fluency). Using pilot data, and prior to primary data collection, I considered several solutions drawing from a range of speech rubrics, but, in consultation with four graduate students enrolled in a language assessment course (the same population intended to utilize the rubric), I determined that any rubric would require a change in task procedure (specifically, more in-depth instructions). To ensure comparability across studies, I chose not to collect task assessment for the Picture task, leaving only the Experiential and Academic tasks for this section of the analysis.

Interactive. The four Raters utilized a rubric previously established in Ockey (2009, 2011), and subsequently used in Leaper and Riazi (2014). Used with permission from Dr. Gary Ockey, the rubric originated as a measure of oral group performance at Kanda University of International Studies (Japan). The rubric (available in Appendix J, as presented in Ockey, 2011 [in Language Learning]), consists of five categories (pronunciation, fluency, grammar, vocabulary, communicative skills/strategies) scored along five 0-4 bands (which allowed for half points to be assigned) accompanied by descriptors. As with Experiential, I chose to assign Speakers an averaged overall score. For Speaker #3, who received scores of 2.5 (pronunciation), 2.5 (fluency), 2.5 (grammar), 2.5 (vocabulary), and 3 (communicative skills/strategies), their overall score was 2.6 (summed score [13] divided by number of categories [5]).

As the original design featured several descriptors directly relevant to the target Japanese population of the initial studies, I rewrote them as follows: “Japanese katakana-like phonology

and rhythm" as "non-nativelike phonology and rhythm", "somewhat katakana-like pronunciation" as "somewhat non-nativelike pronunciation", and "phrases taught in junior high school and beginning high school" as "phrases taught in early language learning contexts." Though Ockey (2009, 2011) had raters assess interactions live, this study followed May (2011) by providing video recordings of each interaction.

Data Analysis

Not all participants provided a speech utterance for all tasks. For the Picture task, a technical issue led to one Japanese Speaker's utterance not being recorded. Due to difficulty completing the Academic task, two Japanese Speakers were unable to produce enough language to fill a 30-second sample. No issues arose for the Experiential task. This left 28 Picture utterances, 29 Experiential utterances, and 27 Academic utterances for analyses.

For the Interactive task, as might be expected, several scheduled Speakers did not attend their session, leaving multiple Speakers without a speaking partner. While it was possible to reschedule by pairing some of these Speakers together to allow for complete sessions, five Speakers were left without speaking partners (all Chinese). In addition, two Japanese Speakers interacted with speaking partners not from the two L1s of focus (see Footnote #1), and I thus removed them from interactive analyses. In total, 22 Speakers (Japanese = 12, Chinese = 10) completed the Interactive task, comprising five Japanese-Japanese, four Chinese-Chinese, and two Japanese-Chinese dyads.

Before coding and analyzing, I transcribed entire task utterances (monologic and interactive). A second transcriber then verified my transcriptions. I then edited down each utterance to the 30- and 60-second excerpts used for speech rating.

Linguistic coding. I coded each speech sample for a series of phonological and fluency measures following guidelines established in Isaacs and Trofimovich (2012) and Trofimovich and Isaacs (2012). Although evidence exists that listeners' subjective rating of linguistic measures aligns strongly with the linguistic coding of similar measures by trained coders (Saito et al., 2017), the current study adopted the latter approach for two reasons. First, a subjective approach (as used in Crowther et al., 2015a, 2015b) required a significant level of commitment on the part of listeners (four 2-hour sessions). With minimal difference between subjective ratings and linguistic coding, I deemed the latter process to be more time efficient.

The second reason to pursue the linguistic coding approach was that it provided a more minute understanding of what and how speech is perceived. A limitation of the subjective approach is that not only was it time intensive, but the complexity of the linguistic measures made it difficult to identify the more minute components of L2 speech. For example, Saito et al. (2017) reduced the 19 measures employed in Isaacs and Trofimovich (2012) to 10, with several measures of fluency (articulation rate, mean length of run, number of filled and unfilled pauses) conflated into a single 'Speech Rate' category, while a single 'Vowel/Consonant Errors' category was devised consisting of both segmental errors and syllable structure errors (additional reduction occurred for lexical, grammatical, and discourse measures as well). To allow for as much in-depth linguistic analysis as possible, the current study employed 11 phonological[12] and fluency measures previously identified in Isaacs and Trofimovich (2012), following the same

[12] I removed one measure of phonology (Pitch Range) as it was not possible to calculate (using Praat) for interactive speech due to overlap of voices. In Isaacs & Trofimovich (2012) and Trofimovich & Isaacs (2012), this measure provided minimal association with both accentedness and comprehensibility ($r <$ .10).

coding guidelines.[13] Table 4 presents the 11 measures, while Appendix K provides the coding guidelines. For readability, I have revised the name of several measures, as provided in Table 4.

Table 4

List of 11 phonological and fluency measures (drawn from Isaacs & Trofimovich, 2012).

Original Name	Revised Name
Segmental Error %	Segmental Accuracy
Syllable structure errors %	Syllable Structure Accuracy
Word stress errors %	Word Stress Accuracy
Rhythm (vowel reduction ratio) %	Rhythm
Pitch contour (intonation error rate) %	Intonation
Filled Pauses	Filled Pauses
Unfilled pauses	Unfilled Pauses
Pause errors %	Pause Appropriateness
Repetitions/self-corrections %	Repetitions/Self Corrections
Articulation rate	Articulation Rate
Mean length of run	Mean Length of Run

Reliability. Reliability measures were calculated for monologic and interactive speech rating, linguistic coding, and task scoring.

Monologic speech rating. Following common conventions in speech production research (Munro & Derwing, 2015), I determined Listener reliability by calculating intraclass correlation

[13] Isaacs & Trofimovich (2012) and Trofimovich & Isaacs (2012) included grammatical, lexical, and discourse measures. As my interest lies in pronunciation training, I here emphasize only the phonological and fluency measures.

coefficients (ICCs) for accentedness and comprehensibility per Speaker per monologic task. As reported in Table 5 reliability was within acceptable levels (> .80; Larson-Hall, 2009). As such, I subsequently calculated a mean score that averaged speech ratings across Listeners for each Speaker per task. These mean scores serve as the accentedness and comprehensibility data for statistical analyses.

Table 5

Intraclass correlation coefficients for accentedness and comprehensibility.

	Picture	Experiential	Academic	Interaction
Accentedness	.879	.888	.919	.948[1]
Comprehensibility	.932	.907	.966	.956[1]

Note. 1 = only the 22 Speakers who completed the interactive task were included.

Interactive speech rating. Interactive speech rating followed the same procedure as above, though only including the 22 Speakers who completed the interactive session of the study. As shown in Table 5, ICCs were within acceptable levels (> .80). However, an additional consideration included the confidence level of Listeners differentiating between the two Speakers in each interaction. Percent of confidence ranged from 59% to 100%, though only one dyad was below 70%. Considering the high reliability between Listeners for accentedness and comprehensibility, I removed only this dyad (mixed L1 Japanese-Chinese) from further analyses, leaving 10 dyads (and 20 Speakers). Combined with the lost monologic samples, only 17 Speakers completed all four tasks.

Linguistic coding. As in Isaacs and Trofimovich (2012), three trained, secondary coders recoded the speech of 12 Speakers (41%) who completed both monologic and interactive tasks. Secondary coders included an undergraduate-level TESOL-minor (Segmental Accuracy, Word

Stress Accuracy, Articulation Rate) and two PhD-level applied linguistic (1: Syllable Structure Accuracy, Rhythm, Intonation; 2: Filled & Unfilled Pauses, Pause Appropriateness, Repetitions/Self-Corrections, Mean Length of Run) students. Table 6 reports ICCs for all categories except Syllable Structure Accuracy. Discussed in more detail at the conclusion of Chapter 4, in short, I removed this category from analysis due to coding concerns raised during discussion with my secondary coder. A second measure, Rhythm, was also removed due to low ICC (.137). This measure involved coding for how accurately Speakers reduced vowel sounds in unstressed syllables and function words, a highly subjective judgment. Despite training, review, and discussion with my secondary coder, we were unable to develop reliability in our coding, with no discernable pattern of differences in perception. Though disappointing considering the high association with both accentedness (r = .74) and comprehensibility (r = .76) in Trofimovich and Isaacs (2012), pursuing Rhythm within the current analysis would not provide much insight given the low reliability.

The remaining categories had ICCs which ranged between .528 and .999. While not ideal for certain categories (Intonation, Pause Appropriateness, Repetitions/Self-Corrections), these measures are also highly subjective, and potentially problematic for coding (as discussed at the conclusion of Chapter 4). Considering the high agreement on the majority of measures, my initial coding was utilized for analyses, though all interpretations of linguistic associations are presented cautiously, given the low ICCs for several variables.

Task scoring. As Raters collaboratively assigned a score per task utterance, there was no need to calculate a measure of reliability.

Analyses parameters. For all analyses provided, alpha is initially set at .05. For linguistic measures, values have been coded so that all positive correlations equate to an increase

in performance, with the exception of Filled Pauses, Unfilled Pauses, and Repetitions/Self-Corrections. Effect sizes follow the guidelines put forth by Plonsky and Oswald (2014)[14] for SLA research, and sample sizes per analysis have been made explicitly clear.

Table 6

Intraclass correlation coefficients for 11 linguistic measures of speech.

	Overall
Segmental Accuracy	.823
Syllable Structure Accuracy	N/A
Word Stress Accuracy	.806
Rhythm	.137
Intonation	.655
Filled Pauses	.938
Unfilled Pauses	.860
Pause Appropriateness	.528
Repetitions/Self Corrections	.610
Articulation Rate	.999
Mean Length of Run	.822

[14] Plonsky & Oswald (2014) proposed the following guidelines for effect sizes in SLA: weak ($r > .25$, $d > .40$), medium ($r > .40$, $d > .70$), and strong ($r > .60$, $d > 1.00$). Due to the use of nonparametric analytic tools, only r is utilized in the below analyses.

CHAPTER 3: RESULTS

Below I report my results in three waves. In Wave 1, I discuss research questions 1-3, which target Listeners' perception of L2 accentedness and comprehensibility across monologic and interactive speech. Next, in Wave 2, I consider research question 4, which focuses on Speakers' knowledge, training, and awareness of phonological measures of L2 speech. Finally, in Wave 3, I address the relationship between Listeners' perception of accentedness and comprehensibility, and these global measures' strength of association with task performance on Experiential, Academic, and Interactive tasks.

Wave 1: Monologic and Interactive Speech Performance

Wave 1 of analyses focused on Listeners' perception of Speakers' accentedness and comprehensibility across monologic and interactive tasks. I included only 20 of 29 Speakers in the current analyses. I removed five Speakers who did not complete both the monologic and interactive tasks, two who did not interact with a Chinese or Japanese speaking partner, and finally one dyad (two Speakers) whom Listeners indicated a limited ability to differentiate between (61% indicated a lack of confidence). In summary, this wave of analysis draws upon 10 dyads comprised of 11 Japanese and 9 Chinese Speakers (9 same-L1, 1 mixed-L1).

Descriptive comparisons. In Table 7, I report mean scores, standard deviations, and 95% Confidence Intervals for accentedness and comprehensibility across task type. For all tasks, Listeners rated Speakers as being easier to understand (comprehensibility) than they were nativelike (accentedness). Listeners found Experiential speech the easiest to understand, and Academic speech the most difficult, while Interactive speech was most nativelike in terms of accentedness, with Academic the least. Figure 4 and Figure 5 present bar graphs depicting accentedness and comprehensibility comparisons both within and between tasks.

Table 7

Speaker performance on monologic + interactive tasks.

		Mean	SD	95% Confidence Intervals	
Picture (N = 19)	accentedness	3.66	0.57	3.39	3.94
	comprehensibility	4.64	0.88	4.22	5.06
Experiential (N = 20)	accentedness	3.83	0.56	3.60	4.15
	comprehensibility	5.23	0.73	4.97	5.70
Academic (N = 18)	accentedness	3.64	0.75	3.27	4.01
	comprehensibility	4.62	1.26	3.99	5.25
Interactive (N = 20)	accentedness	3.84	0.74	3.54	4.28
	comprehensibility	5.02	0.92	4.63	5.56

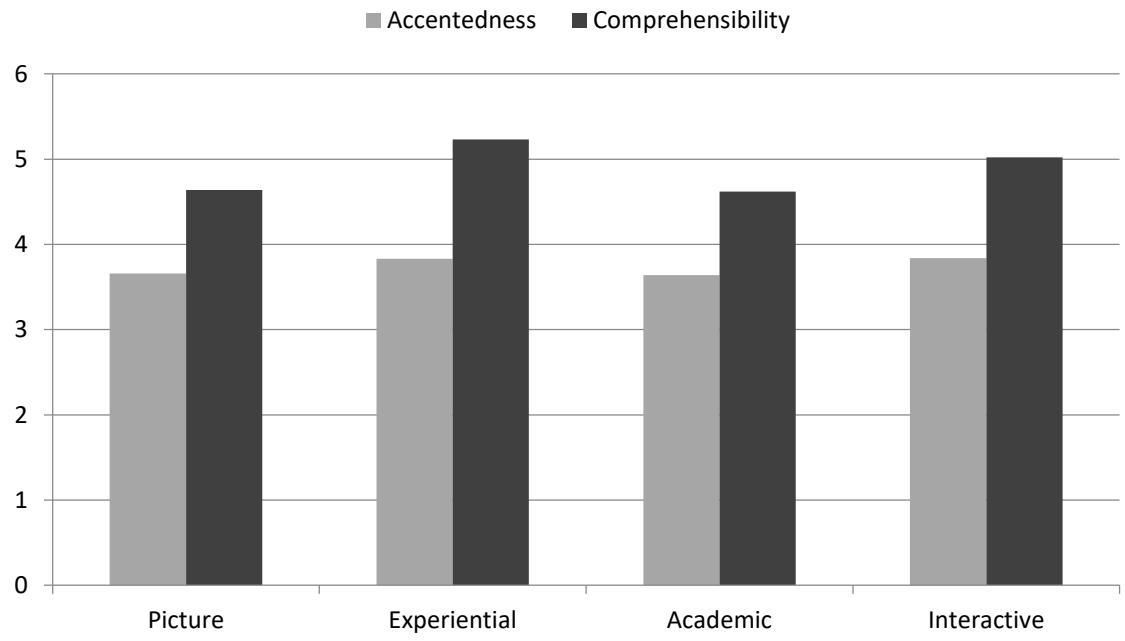

Figure 4. *Comparison of accentedness and comprehensibility ratings within tasks.*

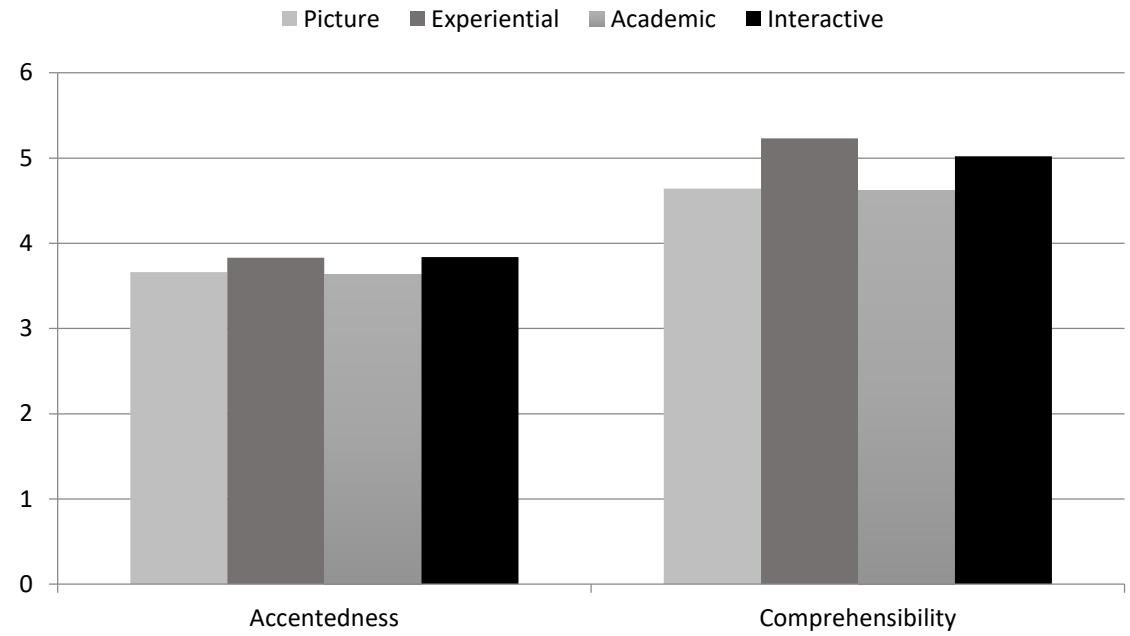

Figure 5. *Comparison of accentedness and comprehensibility ratings across 4 tasks.*

Before investigating whether perceptions of accentedness and comprehensibility across tasks differed significantly, I checked for the existence of a prompt effect within the Experiential, Academic, and Interactive tasks, followed by a review of common assumptions for statistical analyses.

Prompt effect. For Experiential, half the participants completed each prompt (N = 10 each). Considering the small sample size for each prompt, I conducted a nonparametric Mann-Whitney U test to determine if a prompt effect existed.[15] While there was no prompt effect for accentedness (U = 32.500, Z = -1.324, exact p = .190), there was for comprehensibility (U = 16.500, Z = -2.536, exact p = .009). Listeners perceived Speakers as being easier to understand when responding to the Party prompt (M = 5.69, SD = .65) than the Restaurant prompt (M = 4.87, SD = .57). An effect size of r = .56 indicated a medium strength effect. I calculated an

[15] Levene's test returned non-significant (p > .05) values for accentedness and comprehensibility ratings across tasks, indicating no concerns with homogeneity of variance.

effect size for this difference using the equation $r = \frac{Z}{\sqrt{N}}$, where Z represents the z-score returned by the Mann-Whitney U test and N equals the total number of observations.

I ran the same analyses for the two Academic prompts (Social Interaction, Cognitive Dissonance; N = 9 each), and found no prompt effect for either accentedness (U = 39.50, Z = -0.88, exact p = .931) or comprehensibility (U =33.50, Z = -0.62, exact p = .546).

For the Interactive task, which featured three prompts (Activities [N = 4], International Friends [N = 10], Travel [N = 6]), I conducted a Kruskal-Wallis test, which revealed no prompt effect for accentedness (χ^2 = 2.247, p = .325) or comprehensibility (χ^2 = 5.512, p = .064). As the p-value for comprehensibility approached significance, I carried out three separate Mann Whitney U tests to confirm there was no prompt effect, with a manually Bonferroni-adjusted alpha of .017 (α = .05/3). Listeners did not perceive comprehensibility differently between Activities and International Friends prompts (U = 9.000, Z = -1.556, exact p = .142), Activities and Travel prompts (U = 10.000, Z = -0.432, exact p = .762), or International Friends and Travel prompts (U = 10.000, Z = -2.171, exact p = .031).

To summarize, I found a prompt effect only for the Experiential tasks. As the Experiential prompts were drawn directly from official IELTS materials (International English Language Testing System, 2009, 2011), that they elicit different listener perception of comprehensibility is concerning. I will revisit this concern in Wave #3 of analyses.

Tests of parametric assumptions. Recognizing the importance of data exploration prior to conducting parametric analyses (Field, 2009), I first explored the assumption of normal distribution for accentedness and comprehensibility per task.[16] For Picture (N = 19), both the

[16] As the majority of comparisons to be conducted involve comparing speech ratings of the same Speakers (e.g., paired-samples *t*-tests), I assumed homogeneity of variance (Field, 2009; Larson-Hall, 2010).

Kolmogorov-Smirnov test of normality (p = .200 for both accentedness and comprehensibility) and skewness (accentedness = 0.68; COM = 0.88) and kurtosis (accentedness = -0.29; comprehensibility = -0.30) ratios[17] indicated normal distributions (Field, 2009). Figures 6 provides both histograms and boxplots similarly indicating normal distribution.

For Experiential (N = 20), though visual inspection of accentedness (Figure 7) indicates potential concerns for distribution, both the Kolmogorov-Smirnov test of normality (p = .200 for both accentedness and comprehensibility) and skewness (accentedness = 1.03; comprehensibility = -0.63) and kurtosis (accentedness = -0.42; comprehensibility = -0.56) ratios indicate normal distributions.

As with Experiential, while visual inspection of the Academic (N = 18) accentedness histogram (Figure 8) indicates potential concern with distribution, the Kolmogorov-Smirnov test of normality (p = .200 for both accentedness and comprehensibility) and skewness (accentedness = 0.95; comprehensibility = -0.10) and kurtosis (accentedness = -0.78; comprehensibility = -0.79) ratios do not.

Unlike Picture, Experiential, and Academic, Interactive (N = 20) did not demonstrate normal distribution. The Kolmogorov-Smirnov test of normality was significant for both accentedness (p = .024) and comprehensibility (p = .014). While there were no issues based on the kurtosis ratio (accentedness = 1.38; comprehensibility = 0.70), skewness ratios (accentedness = 2.53; comprehensibility = 2.41) were both above the threshold of 1.96 (Field, 2009). A visual inspection (Figure 9) indicates that for both accentedness and comprehensibility, Listeners tended to assign lower ratings to Speakers on the 9-point scale. As revealed in the boxplots,

[17] Skewness and kurtosis ratios are *z*-scores calculated by dividing skewness and kurtosis values (minus the mean of the distribution [0]) by their standard error. Values below -1.96 and above 1.96 are considered indicators of non-normal distribution (Field, 2009).

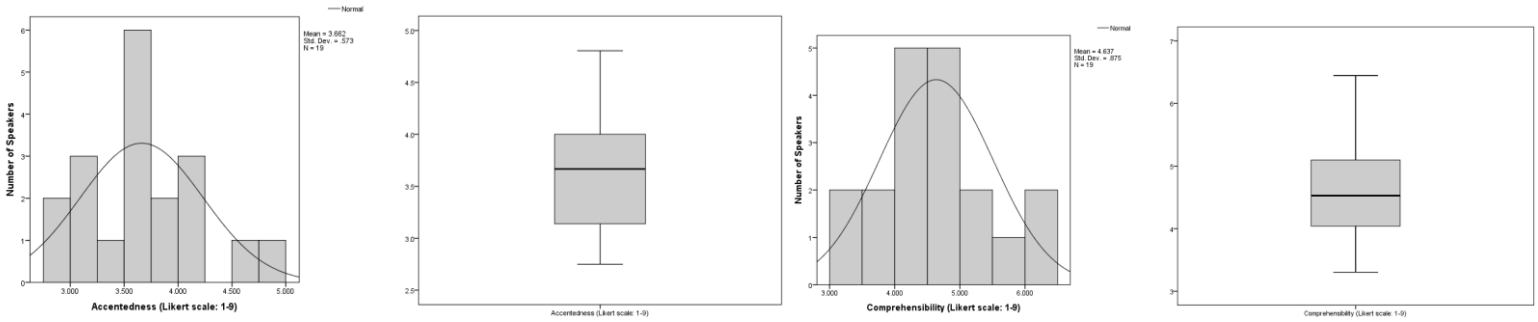

Figure 6. *Histogram and boxplot depicting distribution of accentedness and comprehensibility ratings for Picture task.*

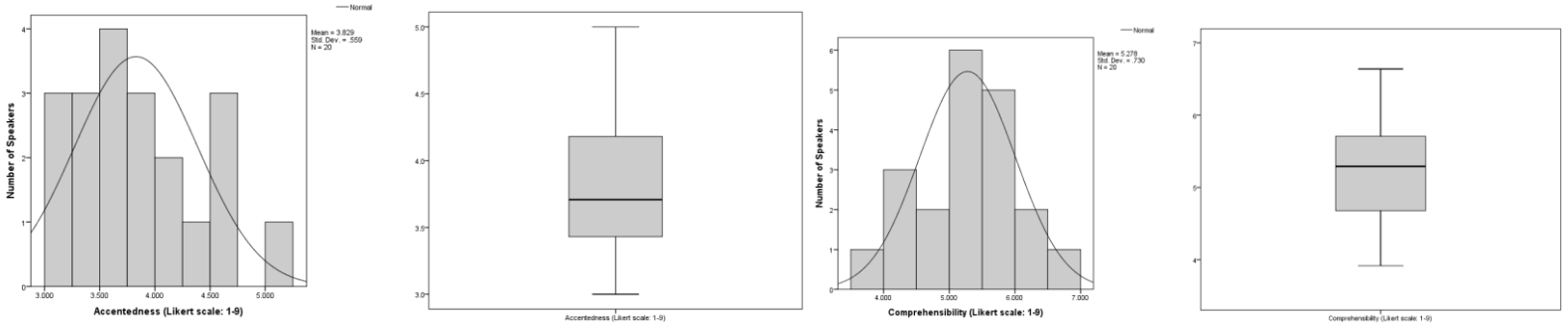

Figure 7. *Histogram and boxplot depicting distribution of accentedness and comprehensibility ratings for Experiential task.*

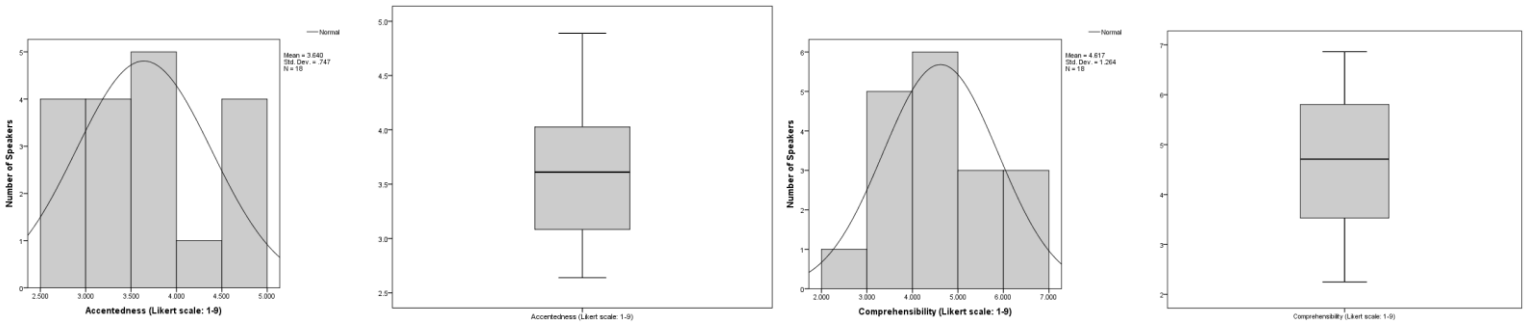

Figure 8. *Histogram and boxplot depicting distribution of accentedness and comprehensibility ratings for Academic task.*

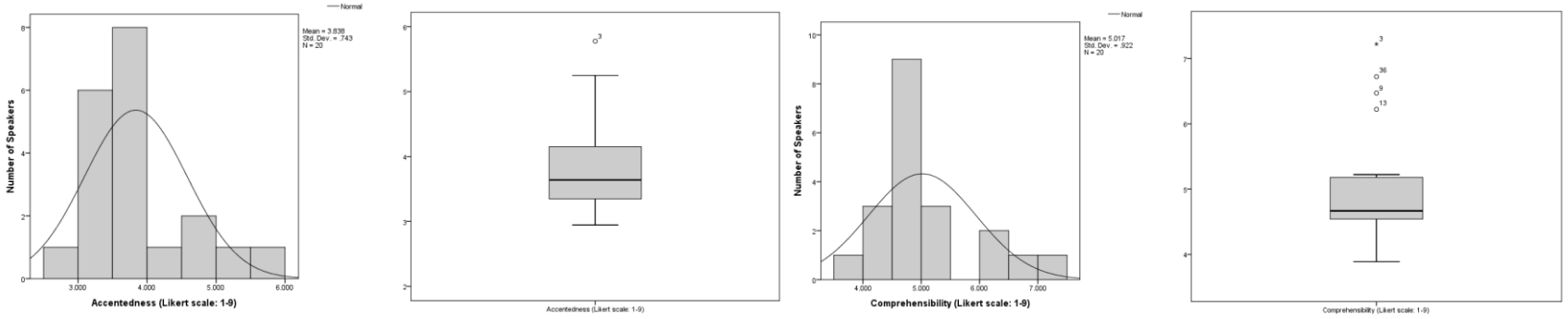

Figure 9. *Histogram and boxplot depicting distribution of accentedness and comprehensibility ratings for Interactive task.*

Speakers #3 and #36 (both Japanese) and #9 and #13 (both Chinese) were positive outliers (though only #3 for accentedness).

The lack of normal distribution found within the Interactive task presents issues when it comes to statistical analyses. One solution would be to remove the four outliers. However, this would reduce the already limited sample size of the study (down to 16, and this is without removing their interactive partners). In addition, outlier removal should be based on the assumption that the observation is not of the population of interest (Field, 2009). However, as all Speakers were drawn from the same IEP environment, this is clearly not the case. While another option would be to attempt to transform the data (Field, 2009), considering the already limited sample size, I chose to instead draw upon nonparametric approaches to data analyses.

Nonparametric analysis. Nonparametric tests can be utilized when data, such as those presented above, do not adhere to the assumption of normal distribution, and instead follow a rank-order system (Field, 2009; Larson-Hall, 2010). In the current analyses, I utilize Spearman's rank tests (in place of Pearson correlations), Mann-Whitney U tests (in place of independent-samples t-tests), Friedman tests (as opposed to one-way repeated measures ANOVAs), and Wilcoxon signed-ranks tests (serving as a substitute for post-hoc paired-samples t-tests).

Accentedness & comprehensibility strength of association. In the first analyses, I utilized Spearman's rank tests to calculate the strength of association between Listeners' perception of accentedness and comprehensibility across tasks. As shown in Table 8, the correlation between the two global measures is quite strong (> .60). This indicates that as accentedness ratings increase (i.e., Listeners deem Speakers to be more nativelike), so do comprehensibility ratings (i.e., Speakers are easier to understand).

Table 8

Spearman's rank (ρ) test results for accentedness and comprehensibility across 4 tasks.

Picture	Experiential	Academic	Interactive
(N = 19)	(N = 20)	(N = 18)	(N = 20)
.756	.789	.899	.665

Accentedness & comprehensibility group differences. To check for group differences between Listeners' perception of Chinese and Japanese Speakers' accentedness and comprehensibility, I conducted a series of Mann-Whitney U tests across tasks.[18] I included a manually Bonferroni-adjusted alpha of .006 (.05/8). I found no significant differences between group ratings, as shown in Table 9. As such, I conducted subsequent analyses on the entire sample.

Table 9

Mann-Whitney U test results for group differences in accentedness and comprehensibility across 4 tasks.

	Picture		Experiential		Academic		Interaction	
	ACC	COM	ACC	COM	ACC	COM	ACC	COM
U	38.00	29.00	33.50	28.00	36.00	34.00	28.50	35.00
Z	-0.77	-1.47	-1.12	-1.55	-0.93	-1.08	-1.51	-1.00
Exact *p*-value	.473	.157	.270	.135	.384	.305	.135	.343

Accentedness & comprehensibility within task comparisons. As presented earlier in Table 7, Listeners rated Speakers as being more comprehensibility than they were nativelike in

[18] Levene's test returned non-significant ($p > .05$) values for accentedness and comprehensibility ratings across tasks, indicating no concerns with homogeneity of variance.

their speech. I conducted a series of Wilcoxon signed-ranks tests to determine whether this difference was significant, with a manually Bonferroni-adjusted alpha value of .013 ($\alpha = .05/4$). For all four tasks, the difference was found to be significant ($p \leq .001$), with a quite strong effect ($r > .80$). Table 10 reports full details of this analysis.

Table 10

Results of Wilcoxon signed-ranks tests between accentedness and comprehensibility across 4 tasks.

	Picture	Experiential	Academic	Interaction
N	19	20	18	20
Mean Difference	0.98	1.40	0.98	1.18
z-score	-3.70	-3.93	-3.46	-3.92
p-value	< .001	< .001	.001	< .001
r^1	.85	.88	.82	.88

Notes. N = sample size; 1 = effect size *r* calculated using $r = \frac{Z}{\sqrt{N}}$.

Accentedness between task comparisons. I conducted a Friedman test to determine if Listeners' perception of Speakers' accentedness differed as a function of the speaking task. It is important to note that the Friedman test considered only the 17 Speakers who completed all four tasks (Table 11 reports task means and standard deviations for these 17 Speakers). The test indicated that a significant difference between tasks existed, $\chi^2 = 9.055$, $p = .029$. To determine the source(s) of this difference, I performed six post-hoc Wilcoxon signed-ranks tests, with a manually Bonferonni-adjusted alpha value of .008 ($\alpha = .05/6$). Table 12 reports the full post-hoc results.

Table 11

Mean (SD) performance on monologic + interactive tasks for Friedman test (N = 17).

		Mean	SD	95% Confidence Intervals	
Picture	accentedness	3.67	0.58	3.38	3.97
	comprehensibility	4.70	0.91	4.23	5.16
Experiential	accentedness	3.83	0.53	3.55	4.10
	comprehensibility	5.32	0.76	4.93	5.71
Academic	accentedness	3.57	0.70	3.21	3.93
	comprehensibility	4.53	1.24	3.89	5.16
Interactive	accentedness	3.83	0.68	3.48	4.19
	comprehensibility	5.00	0.86	4.55	5.44

Table 12

Results of Wilcoxon signed-ranks tests comparing accentedness ratings across 4 tasks.

	Mean Difference	*z*-score	*p*-value	r^1
Picture-Experiential	0.16	-0.82	.410	.20
Picture-Academic	0.10	-0.29	.776	.07
Picture-Interactive	0.16	-0.63	.528	.15
Experiential-Academic	0.26	-1.83	.067	.44
Experiential-Interactive	0.00	-0.08	.940	.02
Academic-Interactive	0.26	-2.39	.017	.58

Notes. 1 = effect size *r* calculated using $r = \frac{Z}{\sqrt{N}}$.

While no significant differences were found, comparisons between Academic and Experiential (p = .067) and Academic and Interactive (p = .017) could be argued to be approaching significance. An investigation into the effect of these two differences revealed a medium strength effect (Academic-Experiential = .44; Academic-Interactive = .58), where Listeners rated Speakers as being more nativelike on both Experiential and Interactive than they were on Academic.

Comprehensibility between task comparisons. I conducted the same analyses described above for comprehensibility between tasks. Again, the Friedman test indicated that a significant difference existed, $\chi^2 = 8.432$, p = .038. Post-hoc Wilcoxon signed-ranks tests, again with a corrected alpha of .008, indicated no significant differences, although comparisons between Experiential and Picture (p = .009), Experiential and Academic (p = .020), and Interactive and Academic (p = .026) were all approaching significance. For the Picture-Experiential comparison, the strength of this effect was strong (r = .63), as Listeners rated the Experiential speech as easier to understand, while for the Experiential-Academic (r = 57) and Interactive-Academic (r = .54) comparisons this effect was medium. In both cases, Listeners rated the Academic speech as more difficult to understand. In addition, a consideration of the Picture-Interactive (r = .38) and Experiential-Interactive (r = .35) comparisons reveal a weaker, but present, effect, with Interactive speech perceived as easier to understand than Picture, but more difficult than Experiential. Table 13 reports full results of the six post-hoc tests.

Spearman correlations. In the final analysis, I calculated Spearman correlations between the nine coded linguistic measures of speech and Listeners' perception of accentedness and comprehensibility.

Table 13

Results of Wilcoxon signed-ranks tests comparing comprehensibility ratings across 4 tasks.

	Mean Difference	*z*-score	*p*-value	r^1
Picture-Experiential	0.62	-2.60	.009	.63
Picture-Academic	0.17	-0.36	.722	.09
Picture-Interactive	0.30	-1.57	.117	.38
Experiential-Academic	0.79	-2.33	.020	.57
Experiential-Interactive	0.32	-1.46	.145	.35
Academic-Interactive	0.47	-2.22	.026	.54

Notes. 1 = effect size *r* calculated using $r = \frac{Z}{\sqrt{N}}$.

Accentedness. Table 14 reports Spearman's rho (ρ) coefficients for accentedness. For clarity of reading, Table 15 summarizes the associations based on strength. For each task, Listener perception of accentedness revealed different patterns of associations. For Picture, Pause Appropriateness has the strongest association. For Experiential, Segmental Accuracy has the strongest association. Measures of fluency (Articulation Rate, Mean Length of Run) had the strongest influence on Listeners' perception of Academic speech, along with Intonation. Finally, Interactive speech was similar to Academic in the associations with Articulation Rate (the only association across tasks > .60). In addition, each task indicates a series of weaker associations with various measures, with Academic being the most diverse. Only two tasks (Picture, Experiential) reveal an association with Segmental Accuracy.

Table 14

Spearman's rho (ρ) coefficients between accentedness and 9 linguistic measures of speech.

	Picture (N =19)	Experiential (N = 20)	Academic (N = 18)	Interactive (N = 20)
Segmental Accuracy	.36	.42	.14	.16
Word Stress Accuracy	-.19	-.03	-.06	-.01
Intonation	.24	-.04	.41	.05
Filled Pauses	-.24	-.09	.01	-.05
Unfilled Pauses	.05	-.01	-.14	-.05
Pause Appropriateness	.45	.21	.36	-.01
Repetitions/Self Corrections	-.18	-.27	-.36	-.04
Articulation Rate	.25	-.04	.48	.60
Mean Length of Run	.21	.01	.40	.30

Notes. > .60 = Strong, > .40 = Medium, > .25 = Weak.

Comprehensibility. Spearman correlation coefficients are presented in Table 16, and a summary of association strength is presented in Table 17. Unlike accentedness where the four tasks tended to demonstrate different patterns, there appears to be more alignment for comprehensibility. All four tasks show associations with two measures of fluency (Articulation Rate, Mean Length of Run). However, these associations are stronger for Picture and Academic than they are for Experiential and Interactive (with Interactive sitting in the middle). In addition, the three monologic tasks feature associations with Pause Appropriateness (though of different strength), while both Experiential and Academic indicate a medium strength association with Repetitions/Self-Corrections. While all four tasks demonstrate some associations with phonological measures, these are all of weaker strength, and only Experiential has an association with Segmental Accuracy.

Table 15

Summary of Spearman correlations with accentedness per task type.

	Weak (r > .25)	Medium (r > .40)	Strong (r > .60)
Picture	Segmental Accuracy, Articulation Rate	Pause Appropriateness	
Experiential	Repetitions/Self-Corrections %	Segmental Accuracy	
Academic	Pause Appropriateness, Repetitions/Self-Corrections	Intonation, Articulation Rate, Mean Length of Run	
Interactive	Mean Length of Run		Articulation Rate

Table 16

Spearman's rho (ρ) coefficients between comprehensibility and 9 linguistic measures of speech.

	Picture (N =19)	Experiential (N = 20)	Academic (N = 18)	Interactive (N = 20)
Segmental Accuracy	.12	.29	.20	.10
Word Stress Accuracy	-.25	-.08	-.06	-.24
Intonation	.33	.07	.38	-.01
Filled Pauses	.10	-.22	-.14	-.17
Unfilled Pauses	-.40	-.01	-.13	.06
Pause Appropriateness	.48	.35	.62	-.03
Repetitions/Self Corrections	-.07	-.43	-.42	.03
Articulation Rate	.67	.32	.68	.48
Mean Length of Run	.62	.30	.60	.46

Notes. > .60 = Strong, >. 40 = Medium, > .25 = Weak.

Cluster analysis. To identify any underlying patterns in Listeners' perception of Speakers' accentedness and comprehensibility across tasks, I conducted a hierarchical cluster analysis (HCA). Cluster analysis is a statistical technique that allows for the classification of cases into a number of groups, or clusters. Those within a group are *similar* in regards to target characteristics but are *unlike* those in the other observed groups (Everitt, 1980; King, 2015). Through an objective mathematical function, cluster analysis minimizes variance within groups while maximizing variance between (King, 2015). HCA, one specific technique of cluster analysis, begins with each case as an individual cluster before combining cases into larger and larger clusters based on distance coefficients (Staples & Biber, 2014). Researchers then make use of several sources to determine the optimal number of clusters, including dendrogram and distance coefficient inspection. As such, it must be noted that HCA involves a level of researcher subjectivity.

Table 17

Summary of Spearman correlations with comprehensibility per task type.

	Weak (r > .25)	Medium (r > .40)	Strong (r > .60)
Picture	Word Stress Accuracy, Intonation	Unfilled Pauses, Pause Appropriateness	Articulation Rate, Mean Length of Run
Experiential	Segmental Accuracy, Pause Appropriateness, Articulation Rate, Mean Length of Run	Repetitions/Self-Corrections	
Academic	Intonation	Repetitions/Self-Corrections	Pause Appropriateness, Articulation Rate, Mean Length of Run
Interactive		Articulation Rate, Mean Length of Run	

In the current analysis, 17 Speakers who completed all four tasks served as clustered variables. Their accentedness and comprehensibility ratings across the four tasks served as grouping variables. Ward's method with squared Euclidean distance served as determiners of cluster distance.[19] After inspecting both the HCA dendrogram (Figure 10) and scree plot (Figure 11),[20] I decided on a 3-cluster solution.[21] I report the descriptive information for each cluster in Table 18.

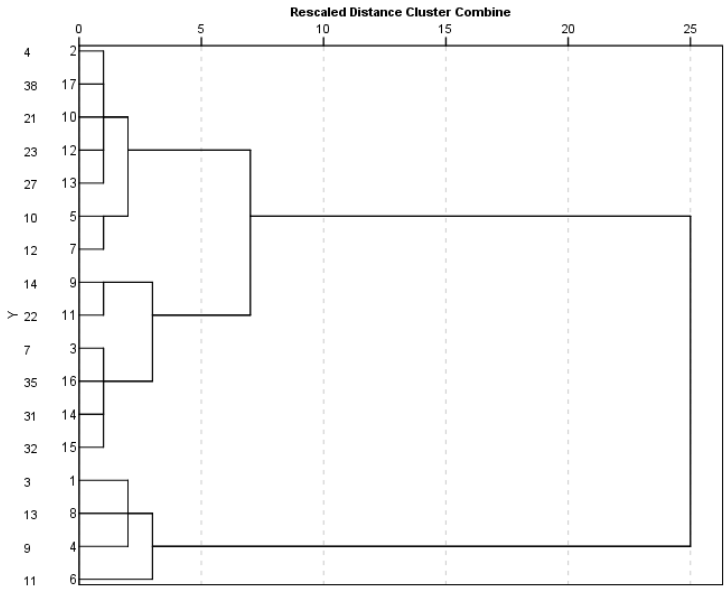

Figure 10. *Dendrogram of hierarchical cluster analysis.*

[19] Ward's method is the most commonly used measure in SLA research (Crowther, Kim, Lee, Lim, & Loewen, forthcoming), and should be paired with squared Euclidean distance (Staples & Biber, 2014).
[20] A scree plot (or approximation of a scree plot) can be created through using an agglomeration schedule. See Staples & Biber (2014) for more details.
[21] When visually inspecting the dendrogram, both a potential 2- and 3-cluster solution existed. Further consideration of the scree plot indicated that the differences in the coefficients begin to flatten out after the third cluster (Staples & Biber, 2014). For this reason, I decided on a 3-cluster solution.

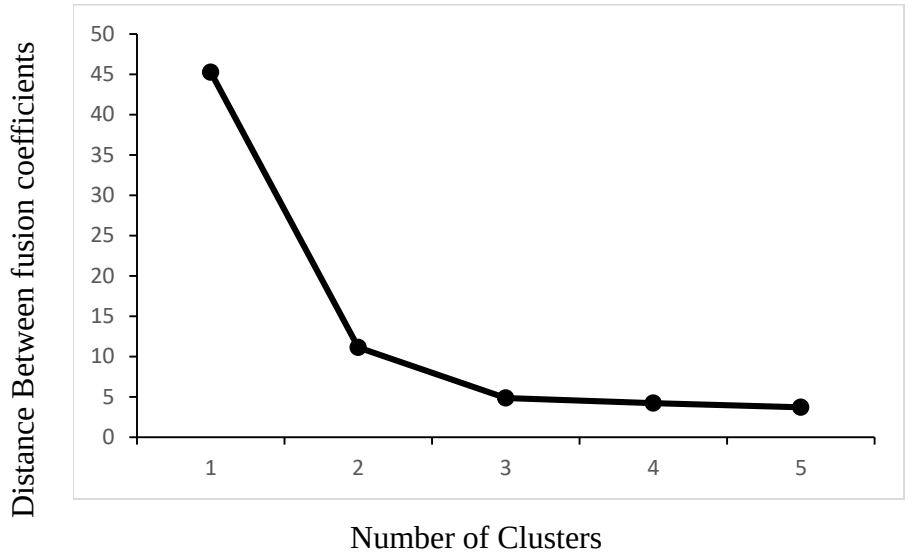

Figure 11. *Scree plot of hierarchical cluster analysis.*

Table 18

Descriptive results for task of 3-cluster solution (mean [SD]).

		Picture		Experiential		Academic		Interactive	
#	N	ACC	COM	ACC	COM	ACC	COM	ACC	COM
1	4	4.40	5.93	4.36	5.98	4.58	6.13	4.84	6.13
High		(0.43)	(0.45)	(0.55)	(0.61)	(0.39)	(0.49)	(0.64)	(1.11)
2	7	3.55	4.53	3.77	5.27	3.58	4.76	3.66	4.85
Middle		(0.24)	(0.57)	(0.47)	(0.72)	(0.20)	(0.44)	(0.24)	(0.27)
3	6	3.34	4.07	3.54	4.94	2.88	3.19	3.36	4.42
Low		(0.57)	(0.63)	(0.37)	(0.69)	(0.21)	(0.50)	(0.23)	(0.32)

Notes. ACC = accentedness, COM = comprehensibility

As a clear continuum exists across tasks for both accentedness and comprehensibility, I have labelled the three clusters as *High*, *Middle*, and *Low*, with Speakers in the High cluster the most nativelike and easiest to understand across tasks and those in the Low cluster the least nativelike and most difficult to understand. Following Staples and Biber (2014), I ran a series of one-way ANOVAs to determine if differences in mean scores between groups was significant.

As seen earlier in the dendrogram, both the Middle and Low clusters branched from the same origin and could be argued to be a single cluster. I thus began by conducting one-way ANOVAs between the High cluster and the combined Middle and Low clusters (I used a manually-adjusted Bonferroni correction of .006 [.05/8]). For all tasks except Experiential, the High cluster was significantly more nativelike and easier to understand ($p \leq .005$). Table 19 provides the full results of this first round of one-way ANOVAs. I then ran the same analysis between the Middle and Low clusters ($\alpha = .006$). Middle only significantly differed from Low on the Academic task ($p < .001$), in which Middle was more nativelike and easier to understand in their speech. Table 20 provides the full results of these one-way ANOVAs.

Table 19

One-way ANOVAs between High cluster and combined Middle/Low clusters.

	Picture		Experiential		Academic		Interactive	
	ACC	COM	ACC	COM	ACC	COM	ACC	COM
F	7.93	13.31	4.07	2.80	52.20	47.81	21.66	10.91
p	.005	.001	.040	.095	< .001	< .001	< .001	.001

Notes. ACC = accentedness, COM = comprehensibility.

Table 20

One-way ANOVAs between Middle and Low clusters.

	Picture		Experiential		Academic		Interactive	
	ACC	COM	ACC	COM	ACC	COM	ACC	COM
F	0.73	1.93	0.94	0.74	37.70	36.39	5.26	7.16
p	.410	.193	.353	.407	< .001	< .001	.043	.022

Notes. ACC = accentedness, COM = comprehensibility.

Recognizing the differences between clusters in task performance, I next investigated the L1 dynamic of each cluster (Table 21). While the High and Middle clusters had members from both L1 backgrounds, all Low cluster members were Japanese. It is important to note that TOEFL (from which the Academic task was derived) serves as a key proficiency examination for non-English speaking international students to gain admission into an English-medium university. As all Chinese Speakers intended to pursue undergraduate study at the university, they were likely more practiced at the task than their Japanese peers, who were exchange students enrolled in IEP.

Table 21

L1 breakdown of 3-cluster HCA solution.

Cluster #	Japanese	Chinese
1	1	3
2	2	5
3	6	0

Wave 2: Participant Patterns

In Wave 2, I considered Speakers' (N = 29) responses to the pronunciation questionnaire. Given the length of the Likert scale used (1-5), my analysis focuses on a descriptive comparison of Japanese (N = 15) and Chinese (N = 14) Speakers' responses. Additionally, although the questionnaire included a category for Speech Rate, I have chosen to remove this variable and

focus solely on the four phonological-based measures. I make this distinction in line with the linguistic measure coding utilized in Isaacs and Trofimovich (2012), where fluency measures were not only treated separate from phonological but were more specific in their classification.

Group responses. As provided in Table 22, Speakers' responses favored Word Stress, which scored the highest for all four survey categories (familiarity, instruction, awareness, importance). In contrast, Speakers' scored Rhythm the lowest across categories. For the two segmental categories, Speakers indicated greater familiarity, instruction, awareness, and importance for Vowel over Consonant production. Considering the two different L1 backgrounds of the Speakers, I next considered if differences existed between Japanese (Table 23) and Chinese (Table 24) Speakers.

Table 22

Group pronunciation survey results (N = 29; Mean [SD]).

	Familiarity	Instruction	Awareness	Importance
Consonants	3.02 (1.07)	2.90 (0.98)	2.74 (1.26)	4.14 (0.92)
Vowels	3.31 (1.22)	3.05 (1.07)	3.16 (1.11)	4.38 (0.90)
Word Stress	3.53 (0.93)	3.38 (1.09)	3.68 (0.97)	4.48 (0.82)
Intonation	3.29 (0.95)	3.07 (1.03)	3.17 (0.85)	4.28 (0.53)
Rhythm	2.69 (0.97)	2.28 (0.88)	2.40 (0.86)	3.55 (0.78)

Japanese responses. In line with the overall group ratings, Japanese Speakers assigned the highest rating to Word Stress across categories, and the lowest ratings to Rhythm. Unlike the overall group ratings, however, Japanese Speakers indicated greater familiarity with Consonants

than Vowels, though the fact they also provided higher scores for Vowel instruction, awareness, and importance leads me to interpret this finding with caution.

Table 23

Pronunciation survey results – Japanese (N = 15; Mean [SD]).

	Familiarity	Instruction	Awareness	Importance
Consonants	3.07 (1.10)	2.60 (0.91)	3.07 (1.22)	4.53 (0.64)
Vowels	2.93 (1.22)	2.67 (1.05)	3.27 (1.03)	4.60 (0.51)
Word Stress	3.67 (1.05)	3.47 (1.30)	4.07 (0.83)	4.67 (0.62)
Intonation	3.20 (0.94)	2.80 (1.21)	3.27 (0.96)	4.40 (0.51)
Rhythm	2.47 (0.83)	2.00 (0.76)	2.33 (0.72)	3.67 (0.72)

Chinese responses. Chinese Speakers indicated a slightly different pattern than did the Japanese. While Word Stress again scored highest for awareness and importance, Vowel production scored highest for familiarity and instruction. Interestingly, while Chinese Speakers maintained the pattern of assigning the lowest scores for familiarity, instruction, and importance to Rhythm, they indicated the least amount of awareness of Consonant production.

Wave 3: Task Performance

The final set of analyses considered the potential relationship between Listeners' perception of accentedness and comprehensibility and Speakers' actual task performance. For Experiential, Academic, and Interactive tasks, I planned to run regression analyses with Task Scores as the outcome variables and accentedness and comprehensibility ratings as predictor variables. For all analyses, I included the entire sample that completed each task (Experiential = 29, Academic = 27, Interactive = 20).

Table 24

Pronunciation survey results – Chinese (N = 14; Mean [SD]).

	Familiarity	Instruction	Awareness	Importance
Consonants	2.96 (1.08)	3.21 (0.97)	2.39 (1.24)	3.71 (0.99)
Vowels	3.71 (1.12)	3.46 (0.97)	3.04 (1.22)	4.14 (1.17)
Word Stress	3.39 (0.79)	3.29 (0.85)	3.29 (0.97)	4.29 (0.97)
Intonation	3.39 (0.98)	3.36 (0.74)	3.07 (0.73)	4.14 (0.53)
Rhythm	2.93 (1.07)	2.57 (0.94)	2.46 (1.01)	3.43 (0.85)

Experiential. I began by considering the strength of association between accentedness, comprehensibility, and Overall, Pronunciation, and Fluency scores. There was a medium association between accentedness and Overall score (ρ = .44) and a strong association between accentedness and Pronunciation score (ρ = .62), though no association with Fluency (ρ = .06) score. For comprehensibility, the associations with both Overall (ρ = .62) and Pronunciation (ρ = .66) were strong, and with Fluency weak (ρ = .29). I next ran a series of hierarchical linear regressions. The first was with Speakers' Experiential Overall scores, the second with their specific Pronunciation score, and the last with their Fluency score. In these regressions, I treated Experiential task score as a continuous variable, as it was calculated by first summing a Speaker's score across categories, then dividing this score by the number of categories. Before beginning, I investigated the potential prompt effect identified earlier. As discussed, Speakers were rated as easier to understand on the Party prompt than they were on the Restaurant prompt. To see whether a similar issue existed for task score, I ran a linear regression with Overall score as the outcome variable and Prompt as the predictor variable (Reference = Restaurant). The

prompt difference only predicted 2% of variance in Overall score (R^2 = .023), and thus prompt was not considered further in this analysis.

I next ran a hierarchical linear regression with Overall score as the outcome variable and accentedness and comprehensibility ratings as predictor variables. Although accentedness and comprehensibility correlated highly (ρ = .79) this was still below the .80 threshold put forth by Field (2009) for multicollinearity. In line with Levis' (2005) Intelligibility Principle, comprehensibility was placed into the model second, to investigate whether comprehensibility explained any variance beyond that explained by accentedness. Table 25 provides the results of the regression. Accentedness and comprehensibility accounted for 33% of total variance in Overall score, with comprehensibility contributing an additional 13% beyond that of accentedness (p = .023).

Table 25

Experiential hierarchical regression results for Overall score.

		B (SE)	β	Adj. R^2	R^2 Change	*p*
Model 1	Constant	2.33 (1.00)				
	accentedness	0.76 (0.27)	.48	.20	.20	.008
Model 2	Constant	1.32 (1.01)				
	accentedness	0.07 (0.37)	.04			
	comprehensibility	0.69 (0.28)	.58	.33	.13	.023

Following Field (2009) and Larson-Hall (2010), I completed my analysis by reviewing the additional assumptions of linear regression analyses. Positively, I found no concerns with multicollinearity (VIF = 2.36, Tolerance = .42), assumption of independent errors (Durbin-Watson = 1.68) or outliers (residual statistics between -3.0 and 3.0, Cook's distance < 1.0, and

Mahalanobis distance < 11). However, an analysis of the P-P and Residual plots indicate slight concerns for the distribution of residuals and homogeneity of variances respectively. In Figure 12, it is clear that there is slight deviation (curvature) from what would be a normal distribution of residuals in the P-P plot and a slight restriction of data points towards the lower middle and left side of the Residual-scatter plot. Neither deviation would appear to be severe, though I consider my interpretations with slight caution.

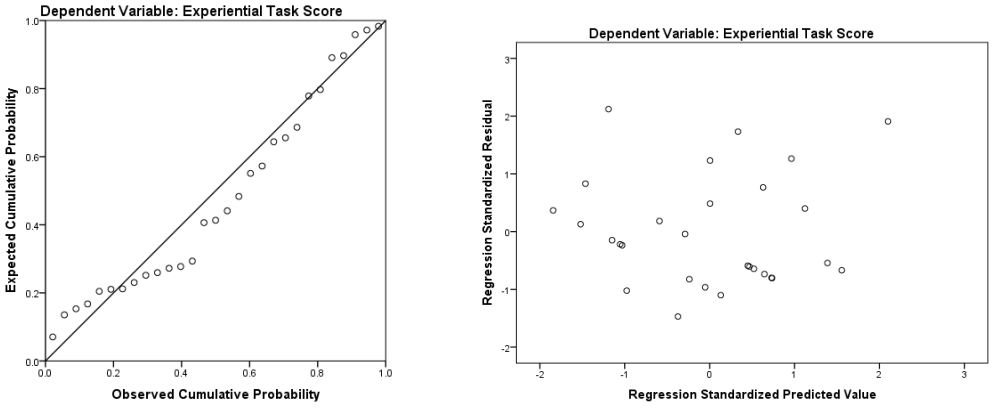

Figure 12. *P-P and residual-scatter plots for Experiential hierarchical linear regression.*

To investigate Experiential performance more closely, I ran a second regression with Speakers' Pronunciation score as the outcome variable, again with accentedness and comprehensibility as predictor variables. This analysis should be viewed as exploratory as scores within a category are arguably categorical, as placement is based on meeting the requirements of the descriptors provided. However, due to the limited sample size, conducting a multinomial regression was not possible (see Academic results below for greater detail). The linear regression revealed that accentedness and comprehensibility accounted for 45% of variance in Pronunciation score, with comprehensibility providing an additional 7% beyond that of accentedness. I next ran the same analysis for Fluency scores and found that only

comprehensibility was a significant predictor of variance (12%). The full regression results are presented in Table 26 (Pronunciation) and Table 27 (Fluency).

Table 26

Experiential hierarchical regression results for Pronunciation score.

		B (SE)	β	Adj. R^2	R^2 Change	*p*
Model 1	Constant	-0.43 (1.32)				
	accentedness	1.48 (0.35)	.63	.38	.38	< .001
Model 2	Constant	-1.64 (1.36)				
	accentedness	0.66 (0.50)	.28			
	comprehensibility	0.83 (0.38)	.47	.45	.07	.039

Table 27

Experiential hierarchical regression results for Fluency score.

		B (SE)	β	Adj. R^2	R^2 Change	*p*
Model 1	Constant	3.70 (1.54)				
	accentedness	0.37 (0.41)	.17	-.01	-.01	.367
Model 2	Constant	2.26 (1.58)				
	accentedness	-0.61 (0.58)	-.29			
	comprehensibility	0.98 (0.44)	.61	.12	.13	.035

As before, I checked for potential violations of assumptions, with minimal concern identified (see Figure 13 for P-P and Residual-scatter plots).

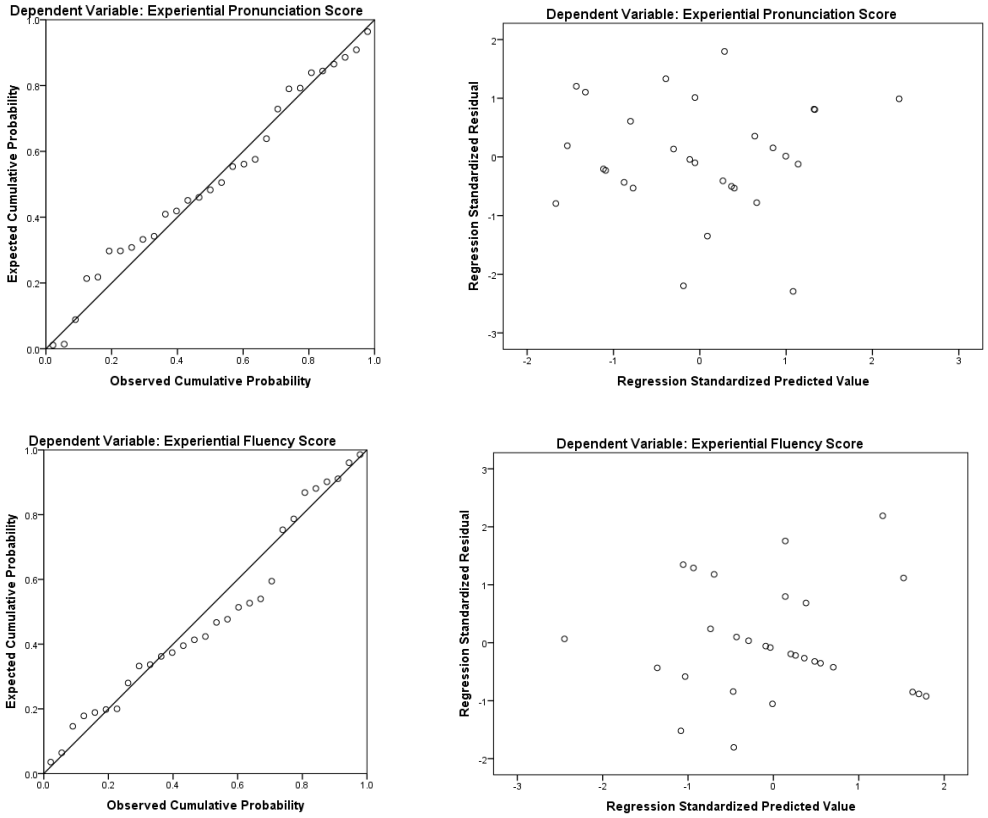

Figure 13. *P-P and residual-scatter plots for Experiential Pronunciation and Fluency hierarchical linear regressions.*

Linguistic associations. I finally checked the strength of association between the nine coded linguistic measures and Experiential Overall, Pronunciation, and Fluency scores. Table 28 presents results. The patterns for Overall and Pronunciation were similar, with Segmental Accuracy being the strongest association for both, with a weaker association with Pause Appropriateness. The only difference was in the weak association for Overall with Articulation Rate. As might be expected for Fluency, there was a strong association with Articulation Rate, Mean Length of Run, and Unfilled Pauses, and weaker associations with Repetitions/Self-Corrections.

Table 28

Spearman's rho (ρ) coefficients between Experiential Overall, Pronunciation, and Fluency scores and 9 linguistic measures of speech (N = 20).

	Overall	Pronunciation	Fluency
Segmental Accuracy	.49	.52	.20
Word Stress Accuracy	-.08	-.06	.03
Intonation	-.00	.04	-.03
Filled Pauses	.17	.16	.12
Unfilled Pauses	-.19	.01	-.47
Pause Appropriateness	.26	.27	.23
Repetitions/Self Corrections	-.21	-.04	-.35
Articulation Rate	.36	.03	.55
Mean Length of Run	.11	-.14	.49

Notes. > .60 = Strong, > .40 = Medium, > .25 = Weak.

Academic. Unlike Experiential, I treated Academic task scores as categorical, as Raters assigned each Speaker one out of five possible scores (0-4). I intended to run a multinomial logistic regression with Task score as the outcome variable (Reference = 1)[22] and accentedness and comprehensibility as continuous predictor variables. However, the correlation between accentedness and comprehensibility was strong (ρ = .920), indicating multicollinearity, so it was not possible to enter both as predictor variables. Again drawing on Levis' (2005) Intelligibility Principle, I chose comprehensibility as a predictor variable to investigate whether this measure of understanding would predict task performance. More concerning than the high correlation between accentedness and comprehensibility was that my limited sample size did not allow me

[22] I treated the '1' band as the reference category, as Raters did not place any Speaker into the '0' band.

to enter a full range of predictors per Band level. As shown in Table 29,[23] not all comprehensibility values appeared per Band level. While descriptively this may be quite informative, for multinomial logistic regression, values of 0 within cells is problematic, often leading to high standard errors (Field, 2009). This was indeed the case, as the standard error of *B* (odds) for the intercept in my model was quite high (> 5).[24]

Table 29

Crosstabulation of comprehensibility ratings with Academic band scores.

Mean comprehensibility	1	2	3	4
2.00-2.99	2	0	0	0
3.00-3.99	3	1	4	0
4.00-4.99	1	2	2	0
5.00-5.99	0	0	3	4
6.00-6.99	0	1	2	1
7.00-7.99	0	0	0	1

Descriptively, the crosstabulation chart offers some interesting observations. No Speakers with a mean comprehensibility score < 5.00 received placement in the highest Academic band. While 92% of Speakers with a mean comprehensibility score > 5.00 placed within the highest two Academic bands, 60% of Speakers with a mean comprehensibility score < 5.00 placed in the lowest two bands. Similarly, comprehensibility was strongly correlated with Academic Band

[23] For ease of reading, I have presented mean comprehensibility scores as being within a range (e.g., 3.00-3.99). Visually, this helps to reduce the number of cells to be considered, while still maintaining the concern of empty cells.

[24] For interest, the model I ran indicated comprehensibility to be a significant predictor of Academic band placement, χ^2 (1,3) = 17.52, p = .001, though only for placement into Band 4 (B = 3.67, SE = 1.37, Exp(B) = 39.21, 95% CIs = 2.68, 572.71, p = .007).

placement (ρ = .62). Though only exploratory, this does indeed indicate that to at least some extent, comprehensibility may predict Academic performance.

Linguistic associations. As with Experiential above, I checked the strength of association between Academic score and the nine linguistic measures, reported in Table 30. While Segmental Accuracy was the only (weak) association with a phonological measure, Academic scores were strongly associated with fluency measures, specifically Articulation Rate and Mean Length of Run.

Table 30

Spearman's rho (ρ) coefficients between Academic Band score and 9 linguistic measures of speech (N = 27).

	Academic
Segmental Accuracy	.34
Word Stress Accuracy	-.09
Intonation	-.03
Filled Pauses	-.24
Unfilled Pauses	-.34
Pause Appropriateness	.55
Repetitions/Self Corrections	-.38
Articulation Rate	.81
Mean Length of Run	.75

Notes. > .60 = Strong, > .40 = Medium, > .25 = Weak.

Interactive. I first considered the strength of association between both accentedness and comprehensibility with Interactive Overall, Pronunciation, and Fluency scores. No significant associations were found (see Table 31). This was echoed in the hierarchical linear regression for Interactive Overall score. As with Experiential, I treated Overall scores as a continuous variable

(summed score divided by total number of categories). Accentedness correlated with comprehensibility at a high but acceptable level (r = .67) and was again entered into the model first. As shown in Table 32, the hierarchical linear regression indicated that accentedness and comprehensibility combined explained a minimal 5% of variance in Interactive task scores (p = .892), with higher comprehensibility even appearing to have a negative impact on Overall score. Similar to Experiential, my inspection of assumptions indicated limited concerns, though the P-P plot features slight deviation from linearity and the residual plot clearly shows a restriction of data points to the lower right side (Figure 14). Considering the minimal association strengths, I did not pursue a regression with either Pronunciation or Fluency scores.

Table 31

Spearman's rho (ρ) coefficients between accentedness, comprehensibility, Interactive Overall, Pronunciation & Fluency scores (N = 20).

	Accentedness	Comprehensibility	Interactive Score	Pronunciation Score	Fluency Score
Accentedness	-	.665	< .01	-.08	-.03
Comprehensibility	-	-	-.11	-.06	-.03

Table 32

Interactive hierarchical regression results (N = 20).

		B (SE)	β	Adj. R^2	R^2 Change	*p*
Model 1	Constant	3.35 (0.50)				
	accentedness	-0.14 (0.13)	-.25	.01	.01	.292
Model 2	Constant	3.37 (0.55)				
	accentedness	-0.10 (0.30)	-.18			
	comprehensibility	-0.33 (0.24)	-.07	-.05	-.06	.892

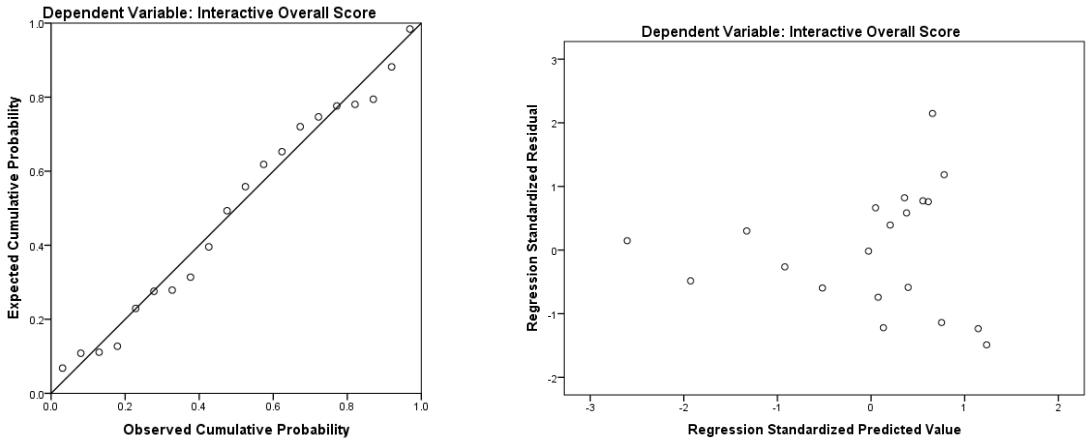

Figure 14. *P-P and residual-scatter plots for Interactive hierarchical linear regression.*

Linguistic associations. Raters' Overall, Pronunciation, and Fluency scoring reveal three different and somewhat unexpected patterns. For both Overall and Pronunciation, it appears that the presence of Filled Pauses (e.g., er, um, uh) led to lower scores, whereas more Unfilled Pauses led to higher scores. For Overall, there was also a weak association with Articulation Rate, though the negative association indicates that less production equaled a higher task score. This may be linked to Raters' expectation that both Speakers would equally contribute, so those who were more dominant may have scored lower. For Pronunciation, one weak phonological association included Intonation, though this association is not at all intuitive (less accurate Intonation led to greater Pronunciation scores). For Fluency, Filled Pauses had the strongest association (less Filled Pauses equated to higher Fluency score), and a weaker association existed with Segmental Accuracy. That Segmental Accuracy was associated with Fluency and not Pronunciation is an unexpected finding.

Table 33

Spearman's rho (ρ) coefficients between Interactive Overall, Pronunciation, Fluency scores and 9 linguistic measures of speech (N =20).

	Overall	Pronunciation	Fluency
Segmental Accuracy	.07	-.15	.32
Word Stress Accuracy	-.02	.15	-.01
Intonation	-.18	-.31	-.14
Filled Pauses	-.27	-.31	-.46
Unfilled Pauses	.27	.49	.18
Pause Appropriateness	-.17	.04	.06
Repetitions/Self Corrections	.08	.16	-.10
Articulation Rate	-.36	-.24	-.08
Mean Length of Run	-.22	-.20	.05

Notes. > .60 = Strong, > .40 = Medium, > .25 = Weak.

CHAPTER 4: DISCUSSION

In this section I discuss my findings in relation to the six research questions posed at the outset of my dissertation. I first provide a summary of the results for each question.

Summary of Research Questions and Findings

Task effect. The first wave of analysis addressed three research questions:

1. Does listener perception of L2 accentedness and comprehensibility differ as a function of task (monologic vs. interactive)?

Though minimal, task type did indeed elicit different Listener perception of accentedness and comprehensibility. Within task, Listeners' perceived Speakers to be easier to understand than they were nativelike in their speech, with a strong effect ($r > .80$) across tasks. Between tasks, Interactive speech appeared to differ only from TOEFL speech, both in terms of perceived accentedness and comprehensibility. It should be noted, though, that this difference only approached significance (albeit with a medium strength effect, $r > .54$).

2. Do the linguistic measures of L2 speech that influence listener perception of L2 accentedness and comprehensibility differ as a function of task (monologic vs. interactive)?

There was indeed a difference between the linguistic measures that influenced Listener perception across tasks. For accentedness, the Interactive task was the only task to feature a strong association (Articulation Rate, $r = .60$), and aligned more closely with Academic than either Picture or Experiential. For comprehensibility, all tasks demonstrated associations with fluency measures (Articulation Rate, Mean Length of Run), though the strength of association was strongest for Picture and Academic, with Interactive speech falling in between Experiential

and Picture/Academic. In general, Interactive speech demonstrated the weakest associations with phonological measures across accentedness and comprehensibility.

3. Do listeners' perception of Speakers' accentedness and comprehensibility follow any patterns across task (monologic and interactive)?

Listeners' perception of accentedness and comprehensibility across tasks placed Speakers into three groups. Group 1, High, was more nativelike and easier to understand than those in Groups 2 and 3 on all tasks except Experiential. Group 2, Middle, was more nativelike and easier to understand than those in Group 3, Low, but only on the Academic task. Considering the L1 Japanese make-up of Group #3, this difference is potentially an effect of task familiarity, as Chinese Speakers are likely more practiced than Japanese when it comes to the Academic task (which was based on the TOEFL iBT integrated speaking task).

Pronunciation awareness. The second wave of analyses considered research question #4.

4. What awareness of phonological measures of L2 speech do learners possess?

As a group, Speakers indicated greater familiarity, instruction, awareness, and importance for Word Stress. For all four categories, Rhythm received the lowest score. In regards to Segments, Vowels scored higher across the four categories than did Consonants. However, whereas the Japanese Speakers aligned closely with the overall group perception, the Chinese did not. Instead, the Chinese Speakers indicated greater familiarity and instruction with Vowel production (though Word Stress was still rated highest for awareness and importance).

Task performance. The final research questions focused on the relationship between Listeners' perception of Speakers' accentedness and comprehensibility and Speakers' actual task performance.

5. Does listeners' perception of L2 accentedness and comprehensibility predict overall task performance?

Perceived accentedness and comprehensibility accounted for 33% of variance in Experiential Overall score, 45% in Experiential Pronunciation score, and 12% in Experiential Fluency score. In all cases, comprehensibility provided variance beyond that of accentedness (13% for Overall, 7% for Pronunciation, 13% for Fluency). While no regression analysis for Academic was possible, there was a strong correlation between Academic Band score and comprehensibility (ρ = .62), and descriptive analysis indicated that a comprehensibility rating > 5 was likely to land a Speaker in Bands 3 or 4 90% of the time, while Speakers with scores < 5 were placed in Bands 1 or 2 60% of the time. Finally, neither accentedness nor comprehensibility appeared to predict performance on the Interactive task.

6. Do linguistic measures associated with listeners' perception of L2 accentedness and comprehensibility align with those associated with raters' task scores on monologic and interactive tasks?

For Experiential, both Listeners and Raters attended to similar measures when assigning comprehensibility and task scores. Though the strength of associations differed, common measures included Segmental Accuracy, Pause Appropriateness, Repetitions/Self-Corrections, Articulation Rate, and Mean Length of Run. The only common measure between accentedness and Overall score was Segmental Accuracy. For Academic, while Listeners attended to suprasegmental measures and Raters to segmental, both appeared to emphasize the range of fluency measures (Pause Appropriateness, Repetitions/Self-Corrections, Articulation Rate, Mean Length of Run). This was similar when comparing both accentedness and comprehensibility

associations to those of task score. For Interactive speech, there appeared to be little overlap between which measures Listeners and Raters attended to.

Listener Perception Across Monologic and Interactive Tasks

Exploring task differences. As stated at the outset, one primary goal of this dissertation was to investigate whether Listeners differed in their perception of L2 accentedness and comprehensibility when rating interactive versus monologic speech. In line with Levis' (2005) Intelligibility Principle, my emphasis is on comprehensibility, or the perceived ease or difficulty of understanding (Derwing & Munro, 2015). My previous research had indicated that increased monologic task complexity led to listeners attending to different linguistic measures of L2 oral production (Crowther et al., 2015a, 2017). In addition, Experiential speech, as elicited through the IELTS long turn task, was deemed to be easier to understand than was speech elicited through a Picture Narrative or TOEFL-inspired (i.e., Academic) integrated speaking task (Crowther et al., 2017). The current findings echo this listener ease to some extent, as again Experiential speech was easier to understand for Listeners than that of Picture (r = .63) and Experiential (r = 57). This interpretation, though, is made based on a medium strength effect size rather than p-values (Plonsky, 2015b). Extending beyond the monologic nature of the above studies, my dissertation asked Listeners to rate two Speakers simultaneously as they engaged in an interactive task. Listener perception of comprehensibility for the Interactive task fell in between the easiest to understand Experiential (r = .35) and the more difficult to understand Picture (r = .38) and Academic (r = 54) tasks, though this effect was strongest in the Interactive-Academic comparison. Though still producing slightly more difficult to understand speech, Interactive-elicited speech appears to align more closely with Experiential-elicited speech than

with Academic-elicited speech. Picture-elicited speech, which could be seen as the least "real world" of the four tasks, falls more towards Academic than either Experiential or Interactive.

In regards to comprehensibility, the four tasks fell along the same linguistic-constraint continuum presented in Chapter 2 (Figure 1). The easiest to understand task, Experiential, is the least constraining, while the most difficult to understand task, Academic, is the most constraining. Interactive and Picture similarly take their expected place on the continuum (Interactive closer to Experiential, Picture closer to Academic). It would appear that as greater linguistic constraints were placed on Speakers the less comprehensible their utterances were perceived to be. Though I will discuss specific linguistic measures in more detail below, I would like to note that for all tasks two measures of fluency, Articulation Rate and Mean Length of Run, associated with Listener perception of comprehensibility. The strength of these associations was quite strong ($r > .60$) for the two more constraining tasks, Picture and Academic. Segalowitz (2010, 2016) describes three ways in which fluency may be perceived:

- *Utterance*: the use of measurable temporal features to characterize the fluidity of observable speech.
- *Cognitive*: the cognitive processes responsible for performing a speech act.
- *Perceived*: the subjective judgments of L2 speakers' oral fluency.

One way to interpret the differences in strength of association between fluency measures and perceived comprehensibility across tasks may be to consider cognitive and utterance fluency in more detail. Segalowitz (2016) lists a number of cognitive processes that help to define cognitive fluency, including "the speed and efficiency of semantic retrieval, the handling of the attention–focusing demands inherent in utterance construction, operations in working memory, among others" (p. 82). The greater linguistic constraints placed on Speakers in the Picture and TOEFL

tasks, in which they are required to utilize specific stimuli to formulate a response, likely creates a greater cognitive load through more complex retrieval processes (see Hilton, 2008, for a discussion on the link between lexical knowledge and spoken L2 fluency). This greater retrieval burden may have negatively impacted Speakers' utterance fluency (i.e., poorer performance on temporal measures), and in turn created greater difficulty for Listeners in understanding. This echoes the views of NNS listeners in Crowther et al. (2017), who primarily referenced fluency measures when performing comprehensibility ratings during a think-aloud protocol (unfortunately, no such protocol was utilized with NSs). As linguistic constraints were lessened in the Interactive and Experiential tasks, allowing Speakers to rely on their full range of lexical and syntactic knowledge, it is likely that there was less of a cognitive burden, which led to more balance across Speakers in utterance fluency, and subsequently weaker associations between temporal measures and Listeners' perception of comprehensibility. However, as no specific measures of cognitive fluency were taken, further research is needed before any concrete claims can be made (see Kahng, 2014, for one example of how such measures may be taken).

While my emphasis here is on comprehensibility, it should be noted that Listener perception of accentedness indicated a similar pattern, with Academic the most accented followed by Picture. However, the two least linguistically constrained tasks, Experiential and Interactive, indicated minimal difference in how accented they were perceived to be (r = .02). Both Experiential (r = .44) and Interactive (r = .58) also indicated medium strength differences when compared to Academic, though differences with Picture were minimal (r < .20). That Picture is more aligned with Experiential and Interactive speech for accentedness, rather than Academic as it was for comprehensibility, may be indicative of the type of speech being elicited. Of the four tasks, Academic is likely the most academically orientated. In the Academic task,

Speakers were required to draw upon multiple language skills (listening, reading, speaking) to formulate a response, as the TOEFL-inspired task was designed to mimic the demands of English-medium academic study (Educational Testing Service, 2017). While no measures of vocabulary usage were taken, Saito et al. (2016) previously highlighted the importance of lexical considerations in listener perception of comprehensibility. Similarly, Crowther et al. (2017) provided initial evidence that as task complexity increases, we begin to see an effect of lexical and grammatical measures on perceived accentedness. As the Picture, Experiential, and Interactive tasks required more casual language (share an experience or opinion, tell a story) than that necessary to complete the Academic task, Speakers may have been perceived to be more nativelike in their speech.

A final note of interest, drawing from the HCA, is the relationship between task complexity and Speakers' perceived accentedness and comprehensibility across tasks. For Picture, Academic, and Interactive, there were significant differences between members in the High cluster and those in the Middle and Low ($p < .005$). However, no such difference existed for Experiential. From a pronunciation instruction perspective, this indicates that more cognitively complex tasks may better serve L2 learners, as nativelike and comprehensible speech on such tasks seems to ensure similar production on less complex tasks (Crowther et al., 2017). Another important consideration is task familiarity. For those not in the High cluster, the only difference came down to speech performance on the Academic task. Those in the Low cluster were all Japanese exchange students who reported a TOEIC rather than a TOEFL score (in contrast to their Chinese peers). As the TOEFL-inspired Academic task was not only the most complicated but likely also the most unfamiliar, Low cluster Speakers may have struggled with the task, which in turn elicited lower Listener perceptions of accentedness and comprehensibility.

Interactive alignment. One concern in the above analysis is that while Listeners provided relatively normal distributions for Picture, Experiential, and Academic ratings, they did not do so for Interactive ratings. Rather, they tended to positively skew ratings, grouping Speakers between 3- 4 for accentedness and 4 -5 for comprehensibility. Interestingly, this may be less a comment on the Listeners, but more so on the interactive processes of the Speakers. In responding to a psycholinguistic focus on individual acts of production or comprehension, Garrod and Pickering (2009) highlight how over the course of an interaction, interlocutors demonstrated interactive alignment at both linguistic and non-linguistic levels, often through emulation. Drawing from Fowler, Brown, Sabadini, and Weihing (2003), Garrod and Pickering describe how phonological and acoustic alignment can be quite rapid. Thus, even within a 60-second excerpt, it may be that interlocutors aligned their speech, which led to a more constrained distribution of Listener ratings. The positive skew is likely the result of potentially high performing Speakers who made alignment difficult. In fact, three of the four outliers identified for comprehensibility on the Interactive task were also placed in the *High* cluster during the HCA analysis, while their partners were placed in the *Middle* cluster.[25] I will temper this interpretation, as it is not clear whether 60 seconds is indeed enough time for interlocutors to align in their linguistic output, although it may serve as a starting point for further investigation.

Task complexity. I highlighted earlier how my interactive task was potentially variable in regards to degree of complexity. How Speakers engaged with the task and prompts likely contributed to how complex the task would be. Due to the homogenous population drawn from (IEP 093 & 094, L1 Japanese & L1 Chinese), with dyads that were primarily shared-L1,

[25] The fourth outlier was not included in the HCA, due to a recording issue with her Picture Narrative, and thus not having data across the four task types. However, her speaking partner was placed in the *Low* cluster.

Robinson's (2005) participant variables that might affect complexity were not a factor. While participation measures such as –one-way flow, –few contributions needed, and –negotiation not needed were present, that there was such strong alignment in Listener perception of Experiential and Interactive speech may indicate that the Interactive task itself was not that complex. To support this view, in their post-interaction questionnaires, many Speakers indicated that what made the interactive task easy was the topic itself.[26] In essence, the Interactive task as designed did not require extensive (causal or intentional) reasoning or perspective taking, and required more opinion sharing than negotiating. Without considering different types of interactive tasks, such as picture difference or consensus tasks (e.g., Loewen & Isbell, 2017), it is not possible to make any overarching declarations on whether listeners perceive monologic speech differently than interactive speech. It may be that while the Interactive task employed here aligned well with the Experiential task, a more complex, academic-based interactive task may align more closely with the Academic task used.

Variation in linguistic associations. As already highlighted, for comprehensibility all four tasks shared an association with fluency measures Articulation Rate and Mean Length of Run. This association was strongest for Picture and Academic, which may be a reflection of the greater linguistic constraints of these tasks, which in turn influenced Speakers' cognitive and utterance fluency (Segalowitz, 2010, 2016). One difference between the three monologic tasks and the Interactive task was a monologic association with Pause Appropriateness, not present for Interactive speech. This might be due to the turn-taking nature of interaction. Pausing can be indicative of several fluency processes, including breakdown, repair, and retrieval speed (Skehan, 2009). While pausing is likely to be more detrimental to listener perception when

[26] Further analyses of the post-interaction questionnaires are not included, as Speakers responses were general brief and perfunctory.

occurring mid-clause rather than end-clause (Davies, 2003), misplaced pauses are also likely to be more salient in monologic speech. Within interaction, runs may be shorter and interlocutors may interject in moments where a pause might have occurred (Michel, Kuiken, & Vedder, 2007), and thus inappropriate pauses are less frequent. A pause may also indicate a change in turns, and thus listeners may be less able to attribute a given pause to either interlocutor. From a coding perspective, the inability to assign a between-turn pause to a specific Speaker led to only within-turn pauses being coded. This may help to explain this monologic versus interactive difference.

The final consideration, and directly linked to the primary goal of this study comes down to the phonological associations of monologic versus interactive speech in regards to comprehensibility; specifically, the relevance of suprasegmentals to understanding within an interactive encounter. However, the results of the current study are inconclusive. Based on previous findings it was expected that there would be at minimum an association between Segmental Accuracy and comprehensibility across tasks. Instead, Segmental Accuracy weakly associated only with Listeners' perception of Experiential comprehensibility (r = .29). While several suprasegmental measures were associated across the four tasks, these associations were weak (r < .40), and at times unexpected. For example, lower Word Stress Accuracy appears to elicit higher comprehensibility judgments for the Picture task. While this would seem counterintuitive, a pedagogical emphasis on English stress timing is not unanimously promoted (Low, 2015). While SLA-orientated researchers advocate the need for accurate word stress and rhythm to produce understandable speech (e.g., Benrabah, 1997; Saito & Saito, 2016; Trofimovich & Isaacs, 2012), many ELF-orientated researchers do not (e.g., Deterding, 2010; Jenkins, 2000). Considering the relatively weak association found here, and only for the Picture

Narrative, I am hesitant to comment in depth on this debate. Returning to the general lack of associations with segmental and suprasegmental measures, I propose two potential explanations.

Proficiency consideration. Speakers were IEP students, a designation that entails they are not yet proficient enough in their English ability to pursue full-time undergraduate study. This differs greatly from the communities that made up the samples in Isaacs and Trofimovich (2012) and Crowther et al. (2015a, 2015b, 2017), both of which advocated for a combined emphasis on segmental/suprasegmental measures. In Isaacs and Trofimovich, speakers represented a combined range of proficiencies from beginner to advanced. In Crowther et al., speakers were either undergraduate or graduate students. The findings of Saito et al. (2016) may help to explain the differences in my data compared to those of these similar studies. Saito et al. proposed that optimal rate of speech and adequate and varied prosody were relevant to beginner-intermediate level learners, whereas segmental accuracy and good prosody became relevant only at the advanced stage. Despite the authors determining proficiency based on comprehensibility scores rather than established measures of L2 proficiency, a comparison can still be drawn. If we consider the range of comprehensibility ratings in the current study (mean scores across tasks = 4.62–5.23), this falls within the range of beginner-to-intermediate profiles of Saito et al. (means scores = ~4-6).[27] Linguistic associates of listeners' perception of comprehensibility may thus change in parallel with speakers' increased proficiency. Additional evidence from Derwing, Rossiter, Munro, and Thomson (2004), who worked with low-proficiency L1 Mandarin/L2 English learners, also indicated a strong relationship between fluency and comprehensibility. My Speakers (IEP) were likely less proficient than those of Crowther et al. (undergraduate/graduate),

[27] Saito et al. (2016) employed end points opposite to those utilized in the current study. The estimate provided is an approximate conversion. Following Saito et al., beginner-to-intermediate mean scores ranged from 6.03-4.06.

which may explain differences in perception between the two sets of listeners, specifically the attention to fluency over phonology in my study.

Listener consideration. Greater attention to fluency measures may also be a result of the Listeners employed. Listeners formed a relatively homogenous group, with an age range of 18-25. All were born, raised, and educated in the American Midwest, and indicated minimal exposure to non-native-English speech. This limited familiarity with L2 speech differs greatly from listeners utilized in previous studies by Isaacs and Trofimovich (2012) and Crowther et al. (2015a, 2015b, 2017). In Isaacs and Trofimovich, listeners were also undergraduate students, but lived in the French/English bilingual city of Montréal, Canada. Thus, even without any formal linguistics training, they would have still been exposed to non-native speech on a daily basis (both English-accented French and French-accented English). It should also be noted that the target of rating was French-accented English as well. For Crowther et al., listeners were MA students in an applied linguistics program, and were experienced L2 English instructors, also living in Montréal, Quebec. Evidence from both SLA (Gass & Varonis, 1982) and L2 assessment (Winke & Gass, 2013; Winke, Gass, & Myford, 2013) scholarship has indicated that familiarity with non-native speech can inform/bias listener perception, often in a positive direction (Saito et al., 2017). It may be that such an effect is present in my data. With limited exposure to non-native-English speech, Listeners possibly lacked the skills necessary to accommodate their receptive ability to unfamiliar patterns of fluency (Gallois, Ogay, & Giles, 2005). This may have usurped a focus on more phonological considerations. It is possible that with increased familiarity, a pattern of associations more aligned with those in the studies highlighted above may emerge. Derwing and Munro (2014) provide suggestions for NS listener training that would

aid their comprehension of L2 speech, including accent perception training, background linguistic information of particular L1s, and communication strategies.

Limitations of the current analyses. The above discussion has already highlighted the potential effect of speaker and listener variables on how monologic speech is perceived compared to interactive. The impact of such variables is clearly in need of further investigation. I here highlight three additional limitations to the current study in regards to how I treated interactive speech.

Monologic bias. As stated at the outset, the current study used what had been a monologic-orientated methodology to analyze interactive speech. I did this with the knowledge that I would only gain knowledge from how an outsider (Listener) perceived an interaction, and limited input from those actually involved (Speakers). While it could be argued that this outsider perspective is similar to interactive studies that base their analyses on discourse analysis (e.g., Jenkins, 2000) or LREs (Loewen & Isbell, 2017), it still loses the insight that approaches such as stimulated recall may provide (e.g., Kennedy et al., 2015). Clearly, the data I present are only one half of a complicated story, and further research is needed that looks to more closely bridge the monologic/interactive methodological divide.

Interactive task complexity. I approached this project with the mindset that interaction was the next step in terms of task complexity, moving beyond the monologic tasks employed in my previous work (Crowther et al., 2015a, 2015b, 2017). However, when considering the strong alignment in Listener perception of the Experiential and Interactive tasks utilized in the current study, it may be better to view Interactive speech as existing on its own continuum of complexity, which may or may not align with that for monologic speech. Rather than seeing it as the next step in task complexity, it may be that monologic and interactive speech both exist along

their own continuum, with the potential that more complex interactive tasks (e.g., picture difference, consensus tasks) may lead to different listener perceptions, such as appears to be the case for monologic tasks.

Interlocutor variables. The dyads formed in this study were relatively homogeneous, which allowed me to control for L2 proficiency and linguacultural differences. However, in our globalized world, contact with a range of nonnative speaking partners is likely a daily occurrence (Appiah, 2006; MacKenzie, 2011). Thus, it becomes necessary to consider how different paired/group dynamics may change the architecture of an interaction, and how this may subsequently impact perception, both globally (accentedness, comprehensibility) and linguistically (phonology, fluency). Such considerations moving forward would consider how paired/group dynamics are formed, and the role of proficiency and culture in this formation.

Storch (2002; SLA perspective) and Galaczi (2008; assessment perspective) have both discussed how different group dynamics subsequently impact how interlocutors engage in the co-construction of meaning. Storch describes four distinct dynamics that may form within an interactive event: collaborative, expert/novice, dominant/dominant, dominant/passive (Galaczi refers to them as collaborative, parallel, asymmetric, and blended, respectively). The amount of language produced by each interlocutor in each pairing varies, with collaborative and expert/novice being the most conducive to language development.

A key consideration in how these interactive dynamics form is interlocutors' L2 proficiency. How L2 learners position themselves, or are positioned, within an interaction based on their L2 proficiency has been shown to play an important role in how their interactive ability manifests, with much of this evidence based on performance when paired with same- and different-proficiency interlocutors. Lazarton and Davis (2008), using collaborative decision

tasks, considered how L2 learners in a paired oral assessment did "being proficient", "being interactive", "being assertive", and "being supportive". Their findings indicated that same proficiency speakers tended to work in greater collaboration than when one speaker was of a higher proficiency than the other. In the latter pairing, whether intentional or not, the higher proficiency speaker often reinforced a less proficient identity in their speaking partner, subsequently impacting performance and assessment. A similar effect was demonstrated in Yule and Macdonald (1990), where in an information exchange map task, more proficient learners performing in the role of sender tended to limit the contribution of their less proficient partner. Yet, similar tasks using such mismatched pairings have also been shown to lead to greater production (in terms of word count) from lower proficiency learners (Davis, 2009; Long & Porter, 1985). While the effects of asymmetric pairings have been raised as a concern for construct validity and fairness within L2 assessment (e.g., May, 2009), placing lower proficiency speakers with those of higher ability may not make a difference in raters' perception of general proficiency (e.g., Csepes, 2009; Nakatsuhara, 2004; Norton, 2005).

A second key consideration, often overlooked in lieu of more linguistic considerations (Scollon et al., 2012), is the role of cultural differences in mis- and non-understanding during interaction between speakers of different linguistic and cultural backgrounds. It has been argued that "culture is constructed in discourse" (Galloway & Rose, 2015, p. 160), which has spurred a call for a greater focus on intercultural awareness as a pedagogical target (Baker, 2015; Byram, 1997; Kumaravadivelu, 2008), a focus that allows learners to recognize how a dialogue between language and culture impacts the interactions they engage in (Leung, 2005). Though a goal of convivial relations across interactions is ideal (Crowther & De Costa, 2017), it is far from practical, considering the numerous power differentials that may exist (Norton, 2013). Primarily

related to the possession of material and symbolic resources, those with greater resources may wield greater power within a given interaction, allowing them to shape how an interaction is constructed.

Speakers' Awareness of L2 Pronunciation Measures

My motivation for the L2 pronunciation survey was to determine whether Speakers possessed the metalinguistic awareness to make reference to suprasegmental measures such as words stress, intonation, and rhythm. I hypothesized that a lack of metalinguistic knowledge may help explain the segmental emphasis during LREs and stimulated recall (e.g., Kennedy et al., 2015; Loewen & Isbell, 2017). Interestingly, Speakers actually indicated greater familiarity with and awareness of word stress rather than segmental production. In addition, intonation was rated almost identically to vowel production, which in turn was rated higher across categories than consonant production. Such a focus on word stress and intonation works against my hypothesis that a lack of metalinguistic awareness led to a greater segmental focus during LREs and stimulated recall, which may, in turn, provides support that it is segmental errors that are of greatest importance in attaining mutual intelligibility in interactive contexts (e.g., Jenkins, 2000).

The emphasis on word stress and intonation instruction is also in contrast to what seems to be characteristic of classroom pronunciation practices, where teachers tend to emphasis a segmental focus. One consideration is that previous surveys of classroom pronunciation practices (e.g., Breitkreutz et al., 2001; Foote et al., 2012, 2013) have been conducted in second language contexts. As the vast majority of Speakers (N = 24) indicated no study abroad prior to arrival, their primary English instruction was in a foreign language context. Teacher cognition studies in foreign language contexts are less common (Baker & Murphy, 2011), with much research drawn from EIL/ELF scholars (e.g., Jenkins, 2000; Sifakis & Sougari, 2005). Unfortunately, such

scholarship generally prioritizes teachers' beliefs in regards to NS versus NNS models of English and is less focused on teachers' awareness of specific linguistic dimensions of L2 speech and how this manifests in the classroom. Why my Japanese and Chinese Speakers indicated greater familiarity and instruction in word stress and intonation remains an open question, in need of further investigation (e.g., speaker and teacher interviews, classroom observation).

Accentedness and Comprehensibility Effects on Task Rating

SLA pronunciation scholars have advocated for a pedagogical emphasis on attaining understandable before nativelike speech (e.g., Derwing & Munro, 2015; Levis, 2005). However, in standardized assessment, the constructs of comprehensibility and accentedness have often been conflated in pronunciation assessment scales (e.g., Harding, 2018; Isaacs et al., 2015), making it unclear whether the linguistic measures associated with listeners' perception of comprehensibility (i.e., potential pedagogical targets) are relevant to raters' scoring of task on high stakes assessment. The current study attempted to bridge this gap.

For Experiential, Listener perception of accentedness and comprehensibility was a significant predictor of not only Pronunciation (Adj. R^2 = .45) and Fluency scores (Adj. R^2 = .12), but also Overall score (Adj. R^2 = .33). In addition, comprehensibility accounted for significantly more variance in Overall (p = .023), Pronunciation (p = .039), and Fluency (p = .035) scores than did accentedness. For all three, increased comprehensibility predicts higher task performance. Potentially concerning is the limited amount of variance accounted for in both Pronunciation and Fluency scores. As Overall score takes into consideration not just Pronunciation and Fluency, but also Lexical Resource and Grammatical Range and Accuracy, it is not surprising that comprehensibility explained only 33% of variance in Overall score. However, it is not clear what accounts for the additional 55% of variance in Pronunciation and

88% of variance in Fluency scores. As I utilized the publically available IELTS speaking rubric, one consideration may be the untrained (officially) nature of the Raters (see below), however, Isaacs et al. (2015) indicated that even trained IELTS raters did not always align in what they attend to when assigning learners to a Pronunciation band score. Another consideration may be, as referenced in Chapter 3, that a linear regression was not the appropriate method of analysis, as Band score may be more representative of a categorical variable than an interval one. To investigate this potential source of concern, however, a larger sample size would be needed.

Similar to Experiential, there appears to be evidence that listener perception of comprehensibility may relate to Academic speaking score. This is supported by a strong correlation between the two (ρ = .62), and a descriptive consideration in which those with comprehensibility scores > 5 were placed into Bands 3 or 4 92% of the time. Similarly, those with scores < 5 placed in Bands 1 or 2 60% of the time. It appears that a comprehensibility rating equal to 5 (the mid-point of the comprehensibility scale) may serve as a cut point for assignment into the two higher or two lower Academic speaking bands (no Speaker was assigned a 0 in this study). The small sample (N = 18) and lack of inferential analysis indicates more investigation is needed before any concrete conclusions can be made.

Unlike Experiential and Academic, there was no association found between accentedness/comprehensibility and Interactive Overall, Pronunciation, and Fluency scores. Simply put, it is likely that when scoring Interactive performance, Raters were more attentive to measures of interactive competence (e.g., turn-taking, topic initiation, discourse extension; May, 2011) than they were the actual speech produced. Other considerations may be that Interactive scoring included a video of the interactive event, which would also allow Raters to attend to physical considerations, such as body language (Ducasse & Brown, 2009) when assigning task

scores. In summary, whereas perceived comprehensibility has potential to serve as a predictor of monologic task performance, it appears limited as a predictor for interactive performance.

Alignment of linguistic associations. As overall comprehensibility aligned with task score for both Experiential and Academic, so did the linguistic measures that Listeners and Raters attended to. Similarly, I found no pattern between which linguistic measures influenced Listeners compared to Raters when rating Interactive speech. Whereas the less complex Experiential task featured associations with both phonological (Segmental Accuracy) and fluency (Pause Appropriateness, Articulation Rate) measures, the more complex Academic task emphasized associations with fluency measures (Pause Appropriateness, Repetitions/Self-Corrections, Articulation Rate, Mean Length of Run). Interestingly, the phonological associations for Listeners on Academic were not found for Raters. This may indicate that for Raters, who are focused more on general proficiency than speech perception (Yan & Ginther, 2018), issues in fluency are more salient than phonological concerns. As discussed previously, the Experiential task was less complex than the Academic task, which may have enabled Speakers to produce more fluent speech. Subsequently, this may have enabled Raters to place a greater emphasis on Segmental Accuracy.

Limitations of the task analyses. I make the above interpretations with caution for two important reasons. First, despite using publically available IELTS and TOEFL rubrics, the Raters employed did not receive official IELTS or TOEFL training, and thus are not representative of how official raters may have assessed Speaker performance on either the Experiential or Academic tasks. Second, in official assessment contexts, such as IELTS and TOEFL, an entire speaking battery is employed to holistically assess a speaker's speaking ability. Here, I utilized only a pair of tasks, each derived from a different standardized assessment (IELTS or TOEFL).

As such, the current study should be seen as exploratory. The findings above would indicate that perceived comprehensibility may indeed predict task performance, which provides support that more controlled research comparing listener versus rater perception is necessary

An additional concern, prevalent in much SLA research (Plonsky, 2013), is the limited sample of the study. Sample size clearly impacted the statistical approaches used throughout the study. In addition, from an assessment perspective, I would ideally like to generalize across a much wider-range of L2 learners. That my Speakers represented only two L1s, Japanese and Chinese, and were of a limited proficiency range, IEP 093 and 094, clearly limits my ability to generalize beyond this population.

Causes for Concern: 11 Linguistic Measures of Speech

For this study, I drew upon 11 phonological and fluency measures used previously in Isaacs and Trofimovich (2012). This was done to allow for comparability. While these authors referenced the surprising nature of having gained ICCs > .90 despite the subjective nature of many of the categories (e.g., Segmental Accuracy, Word Stress Acurracy, Rhythm), these measures' application in my study was less concise, and potentially problematic. For example, Word Stress Accuracy received an ICC of .80, which would fall within the acceptable guidelines put forth by Larson-Hall (2010). Yet, when reviewing coding with my secondary coder it became clear that where we both may have identified the same number of errors within an utterance, the specific errors identified did not always align. We achieved agreement on total number of errors, but not on actual errors. Similar concerns exist for Segmental Accuracy, Syllable Structure Accuracy, Rhythm, and Pause Appropriateness, all of which are highly subjective. For example, Syllable Structure Accuracy was defined as any additional or deleted sound (Isaacs & Trofimovich, 2012,

provided the example of an L1 French speaker of English droping the /h/ in ‘holiday’). Consider the following sentence:

a man with a a green suitcase <u>and</u> a woman uh with a green suitcase too

In this utterance produced by Speaker #3, a member of the High cluster, the /d/ in ‘and’ is not pronounced. This was a point of disagreement between my secondary coder and me, as she indicated that such deletion might be seen as characteristic of native-English speech (see Celce-Murcia et al., 2010, for further discussion on this topic). To account for such deletion instances, a choice was made to consider only errors that altered the syllable count of a word (e.g., ‘birthday’ -> ‘birth-u-day’). This not only significantly lowered the number of syllable structure errors from my initial coding, but also created a category no longer directly comparable with previous studies. As such, I subsequently removed Syllable Structure Accuracy from analyses. Rhythm was also removed, but due to an extremely low level of agreement between coders (ICC = .137). These instances are, of course, concerning, as they raise questions on how well the phonological and fluency coding employed actually reflects how Listeners perceived the L2 speech.

For the current study, I have maintained coding, aside from the Syllable Structure Accuracy and Rhythm, to align with Isaacs and Trofimovich (2012). However, moving forward, I intend to recalculate coding using Cohen’s Kappa. This statistical approach provides a strongly conservative estimate of reliability (Plonsky & Derrick, 2016) and considers agreement on each individually coded item. Such analysis would allow for a better understanding of coder perception, specifically any patterns of disagreement that may exist.

CHAPTER 5: CONCLUSION

In this final chapter, I reflect upon the potential theoretical, pedagogical, and assessment implications of the above findings, before providing suggestions for future research.

Implications

In Chapter 2 I proposed potential theoretical, pedagogical, and assessment implications, which I revisit here.

Pedagogical. At the outset I proposed a link between comprehensibility and the Interaction Approach (Long, 1996). Specifically, I proposed that since L2 learners attend primarily to lexical and grammatical measures during communicative breakdowns, and to a lesser extent segmental issues, then a pedagogical focus on suprasegmental measures in the classroom would provide the necessary attention to such measures that previous research has called for (e.g., Derwing & Munro, 2015; Isaacs & Trofimovich, 2012). This proposal assumed that the reason Speakers did not reference such measures during LREs and stimulated recall was due to a lack of metalinguistic awareness. However, my results indicate a greater emphasis on Listeners' perception of Speakers' fluency than for segmental or suprasegmental production. Such an emphasis for beginner-to-intermediate L2 speakers is not unfounded (Derwing et al., 2004; Saito et al., 2016), but pedagogically troublesome (Thomson, 2018). While it may be possible to raise L2 learners' awareness of appropriate pause placement and how to effectively use filled and unfilled pauses, fluency concerns related to lexical, syntactic, and semantic retrieval (Segalowitz, 2010, 2016) are not as easily targeted. Lee et al. (2014), through a meta-analysis of 86 pronunciation instruction studies, indicated stronger effects of explicit pronunciation instruction for beginning and advanced proficiency students than for intermediate students. This may indicate that once a minimum pronunciation threshold is achieved,

instructors' emphasis may be best placed on developing lexical and grammatical knowledge, which in turn would (ideally) raise L2 learners' fluency. The findings of Nagle (2018) may provide initial support for this hypothesis. Focused on L2 Spanish, Nagle measured the growth of perceived accentedness and comprehensibility over a yearlong period. Set in a communicative-based university-level classroom in the US, instructors indicated limited attention to pronunciation during their lessons. Despite no explicit focus on pronunciation, learners still demonstrated general improvement in how comprehensible they were perceived to be. The Interaction Approach theorizes that L2 development occurs through exposure to input, production of output, and engagement in negotiation of meaning (Gass & Mackey, 2015). As learners appear to attend primarily to lexical and grammatical features during communicative breakdowns (e.g., Kennedy et al., 2015; Loewen & Isbell, 2017), it is possible that the increases in comprehensibility observed by Nagle were a direct reflection of increases in lexical and grammatical awareness/knowledge, which in turn would benefit learners' retrieval processes. Once higher fluency is achieved, targeted pronunciation instruction may again be necessary (e.g., promoting segmental accuracy; Saito et al., 2016). However, further research across proficiency levels is needed before making any concrete conclusions on this claim.

Extending the previous discussion on a communicative-based classroom, the question would be the type of tasks necessary to further develop the linguistic measures associated with producing understandable speech. Crowther et al. (2017) proposed that the use of complex tasks would enable learners to practice a wider range of linguistic dimensions (phonology, fluency, lexicon, grammar). The results of my HCA would indicate the same. For the least complex and linguistically constrained task (Experiential) no differences existed between Speakers for perceived accentedness and comprehensibility. However, as task complexity and linguistic

constraints increased, the High cluster began to outperform Mid and Low (in fact, High performed significantly better on Interactive, Picture, & Academic). In addition, the Mid and Low clusters differed only on the most complex, most constrained task (Academic). These findings would appear to indicate that for perceived accentedness and comprehensibility, Speakers who performed well on the more complex, more constrained tasks also performed well on the less complex, less constrained tasks. Essentially, task complexity served to differentiate between Speakers in this study. Pedagogically, as proposed in Crowther et al. (2017), and following on the findings of Nagle (2018), a communicative-based classroom that emphasizes more complex, more linguistically constrained tasks may serve to benefit L2 learners in regards to increasing their ability to produce understandable L2 speech. Clearly more investigation is needed to gauge the potential of such a pedagogical approach.

Assessment. Levis' (2005) Intelligibility Principle has received strong support from SLA scholars in regards to a holistic pronunciation target (e.g., Derwing & Munro, 2015; Isaacs & Trofimovich, 2012). An emphasis in the L2 classroom should be placed on the linguistic measures relevant to producing understandable rather than nativelike speech. From an assessment perspective, that comprehensibility is often conflated with accentedness (Harding, 2017; Isaacs et al., 2015) in rubric descriptors is concerning, as this may problematize what linguistic measures receive focus in the L2 classroom. Even though L2 speakers can be highly comprehensible even while possessing a heavy accent (Derwing & Munro, 2015), would this heavy accent still negatively impact their speaking score? The exploratory findings of the current study may indicate this is not the case, as for both the IELTS- and TOEFL-inspired tasks (i.e., Experiential and Academic, respectively), Listener perception of comprehensibility seems to associate to at least some extent with Raters' task scoring. If so, then the pedagogical emphasis

on the Intelligibility Principle is well founded, in regards to both L2 pronunciation development and speaking assessment.

One limitation of the above proposal is that neither perceived accentedness nor comprehensibility predicted performance on the Interactive task. This raises questions on the ecological validity of comprehensibility as a predictor of interactive success. Much monologic-based research that has measured perceived comprehensibility has done so primarily through audio-recorded utterances, with no visual representation of the speaker (though see Rubin, 1992, and Kang & Rubin, 2009, for exceptions). Such an approach ignores the multimodal nature of communication, where listeners do not simply rely on linguistic cues, but also visual when determining meaning (Jewitt, 2014). In interaction, such visual cues are frequently available. Nonverbal communication may include gesture, posture, facial expressions, and eye behavior (Hardison, in press; Knapp & Hall, 1992). While the effects of nonverbal cues have been investigated in respect to L2 learners' listening comprehension (e.g., Sueyoshi & Hardison, 2005; Suvorov, 2011, 2015; Wagner, 2007, 2008), the importance of such cues has not been considered in respect to listener evaluation of L2 monologic speech. I approached a potential monologic-interactive divide by applying a monologic methodology on interactive speech. However, without considering the availability of visual cues, it may be that this monologic methodology in itself is limited.

Directions for Future Research

As several potential avenues for future research have been referenced previously, I here highlight three that I feel are necessary to continue the line of inquiry presented.

Interlocutor perception. The manipulation of several variables from the current study would be of interest, including speaker proficiency, listener familiarity, and interactive task

complexity. As the potential impact of such variables was discussed in Chapter 4, I here stress the need to emphasize the perception of the interlocutor and compare whether their within-task perception aligns with that of the outside listener. In the current study I employed a methodology characteristic of monologic speech research. The next step would be comparing how these outside-derived, perception-based judgments align with those that may be expressed through the stimulated recall (Gass & Mackey, 2017) of actual participants. Of particular interest would be how these within-task perceptions may change based on various interlocutors, and whether these changes are accounted for in the outside listeners' perceptions of global and linguistic measures of speech.

Task assessment. For IELTS and TOEFL, there is initial evidence that listener perception of L2 speakers' comprehensibility may help predict task performance. As this evidence draws upon Raters without formal IELTS or TOEFL training, however, it can only be viewed as exploratory. In addition, only a single IELTS- and TOEFL-inspired task were considered, as opposed to the entire battery of tasks utilized to assess speaking. The next logical step then, beyond increasing sample size, would be to compare listener perception of L2 comprehensibility to actual IELTS and TOEFL ratings across a range of speaking tasks. As referenced previously, an association between listener perception and rater scoring would add credence to a pedagogical focus on understandable speech (Derwing & Munro, 2015; Levis, 2005).

Linguistic coding. A "methodological turn" (Byrnes, 2013, p. 825) in SLA research has led to a greater emphasis on the methodological rigor employed in conducting empirical research (e.g., Norris & Ortega, 2000; Plonsky & Gass, 2011). While much of this emphasis has been placed on the proper application of statistical procedures (e.g., Cunning, 2012; Plonsky, 2015a;

Plonsky & Gonulal, 2015; Winke, 2014), clearly such methodological review must also encapsulate the initial coding scheme (see Plonsky, Marsden, Crowther, Gass, & Spinner, forthcoming, for an example focused on SLA judgment task design and usage). Linguistic coding, whether targeting phonological, fluency, grammatical, lexical, or discourse domains, has varied greatly across studies. For example, whereas Isaacs and Trofimovich (2012) featured four measures dedicated to grammar and lexicon measures, Saito et al. (2016) utilized six measures for only lexicon. Comparing Kang (2010) to Kahng (2014) highlights the different ways in which L2 fluency can and has been measured. Segmental production has been measured both perceptually (e.g., Isaacs & Trofimovich, 2012) and acoustically (e.g., Solon, Long, & Gurzynski-Weiss, 2017). Pronunciation scholars have made claims that specific linguistic measures are relevant to the production of understandable L2 speech; however, without a uniform approach to linguistic coding, making comparisons across studies is not possible. While it would be unreasonable to expect all scholars to subscribe to the same coding procedures, it does seem necessary to at least make note of what procedures have been used, their strengths and weaknesses, and the reasons as to why researchers have utilized them. A methodological review of linguistic coding within L2 pronunciation research seems well overdue.

Concluding Thoughts

While there is clearly a need to pursue many of the themes of my dissertation further, as highlighted above, there is also insight that can be drawn. The findings of my dissertation continue to support a pedagogical emphasis on intelligible (i.e., understandable) before nativelike speech (Derwing & Munro, 2015; Levis, 2005), and provide evidence that this emphasis is relevant to both L2 communicative and assessment contexts. While greater clarity in regards to which linguistic measures enable an L2 speaker to produce understandable speech is needed, it is

clear that at both a monologic and interactive level, listener understanding is reliant on more than simply a segmental versus suprasegmental debate. Along with the linguistic measures of interest, it is necessary that we consider the proficiency of speakers, the familiarity of the listeners, and the complexity of the tasks employed.

APPENDICES

APPENDIX A

Picture Narrative (Derwing et al., 2009)

THE SUITCASE STORY

The "Suitcase Story" may be used for research purposes only, provided that the user cites the following source:

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. Studies in Second Language Acquisition, 31(4), 533-557.

APPENDIX B

Experiential Task

Version A (IELTS, 2009)

Describe a party that you enjoyed.

You should say:

whose party it was and what it was celebrating

where the party was held and who went to it

what people did during the party

and explain what you enjoyed about this party

You will have to talk about the topic for 1 to 2 minutes.

You will have 1 minute to think about what you are going to say.

You can make some notes to help you if you wish.

Version B (IELTS, 2011)

Describe a restaurant that you enjoyed going to.

You should say:

where the restaurant was

why you chose this restaurant

what type of food you ate in this restaurant

and explain why you enjoyed eating in this restaurant.

You will have to talk about the topic for 1 to 2 minutes.

You will have 1 minute to think about what you are going to say.

You can make some notes to help you if you wish.

APPENDIX C

Academic Task (Educational Testing Service, 2012)

Version A – Social Interaction

Reading Text

People account for their own behavior differently from how they account for the behavior of others. When observing the behavior of others, we tend to attribute their actions to their character or their personality rather than to external factors. In contrast, we tend to explain our own behavior in terms of situational factors beyond our own control rather than attributing it to our own character. One explanation for this difference is that people are aware of the situational forces affecting them but not of situational forces affecting other people. Thus, when evaluating someone else's behavior, we focus on the person rather than the situation.

Listening Text

So we encounter this in life all the time, but many of us are unaware that we do this…even psychologists who study it…like me. For example, the other day I was at the store and I was getting in line to buy something. But just before I was actually in line, some guy comes out of nowhere and cuts right in front of me. Well, I was really annoyed and thought, "That was rude!" I assumed he was just a selfish, inconsiderate person when, in fact, I had no idea why he cut in line in front of me or whether he even realized he was doing it. Maybe he didn't think I was actually in line yet…But my immediate reaction was to assume he was a selfish or rude person.

Ok. So a few days after that, I was at the store again. Only this time I was in a real hurry–I was late for an important meeting–and I was frustrated that everything was taking so long. And what's worse, all the checkout lines were long, and it seemed like everyone was moving so

slowly. But then I saw a slightly shorter line! But some woman with a lot of stuff to buy was walking toward it, so I basically ran to get there first, before her, and well, I did. Now, I didn't think of myself as a bad or rude person for doing this. I had an important meeting to get to–I was in a hurry, so, you know, I had done nothing wrong.

Version B – Cognitive Dissonance

Reading Text

Individuals sometimes experience a contradiction between their actions and their beliefs–between what they are doing and what they believe they should be doing. These contradictions can cause a kind of mental discomfort know as *cognitive dissonance*. People experiencing cognitive dissonance often do not want to change the way they are acting, so they resolve the contradictory situation in another way: they change their interpretation of the situation in a way that minimizes the contradiction between what they are doing and what they believe they should be doing.

Listening Text

This is a true story–from my own life. In my first year in high school, I was addicted to video games. I played them all the time, and I wasn't studying enough–I was failing chemistry; that was my hardest class. So this was a conflict for me because I wanted a good job when I grew up, and I believed–I knew–that if you want a good career, you gotta do well in school. But…I just couldn't give up video games.

I was completely torn. And my solution was to … to change my perspective. See, the only class I was doing really badly in was chemistry. In the others I was, I was okay. So I asked myself if I wanted to be a chemist when I grew up, and the fact is I didn't. I was pretty sure I

wanted to be a sociologist. So … I told myself my chemistry class didn't matter because sociologists don't really need to know chemistry. In other words, I changed my understanding of what it meant to do well in school. I reinterpreted my situation: I used to think that doing well in school meant doing well in all my classes, but now I decided that succeeding in school meant only doing well in the classes that related directly to my future career.

I eliminated the conflict, at least in my mind.

APPENDIX D

Interactive Prompts

Prompt #1
Agree or disagree with the following statement:

It is important to attend many activities <u>when studying abroad</u>.

- Have you attended any new activities while at MSU (sports game, club activity, etc.)?
- Why or why not?
- What type of activities? Did you enjoy them?
- What are positive reasons to attend a new activity?
- What are negative reasons to attend a new activity?
- What type of activity do you like best? A sports activity? An academic activity? Why?

Use specific examples to express your opinion to your partner. Is your partner's opinion similar or different from yours? Try to convince your partner your opinion is best.

Prompt #2
Agree or disagree with the following statement:

It is important to make international friends <u>when studying abroad</u>.

- Do you want to make international friends while here at MSU? Why or why not?
- What are positive reasons to make international friends?
- What are negative reasons to make international friends?
- Have you made any international friends while here? How did you meet them?
- What do you do with your international friends? What events do you attend?

Use specific examples to express your opinion to your partner. Is your partner's opinion similar or different from yours? Try to convince your partner your opinion is best.

Prompt #3
Agree or disagree with the following statement:

It is important to travel to many places <u>when studying abroad</u>.

- Have you visited anywhere while at MSU?
- Why or why not?
- What places have you visited? Why did you choose these places?
- What was your impression of the places you visited?
- What are positive reasons to travel to many different places?
- What are negative reasons to travel to many different places?

Use specific examples to express your opinion to your partner. Is your partner's opinion similar or different from yours? Try to convince your partner your opinion is best.

APPENDIX E

Pronunciation Questionnaire

Consonants

Individual sounds that are not vowels. For examples, /b/, /d/, /g/, & /s/.

How familiar are you with English consonants?

1 2 3 4 5

1 = Not familiar at all *5 = Very familiar*

How much instruction have you received on how to produce English consonants?

1 2 3 4 5

1 = No instruction at all *5 = A lot of instruction*

When speaking English, how aware are you of how you are producing consonants?

1 2 3 4 5

1 = Not aware at all *5 = Very aware*

How important are consonants in producing English speech that is understandable?

1 2 3 4 5

1 = Not important at all *5 = Very important*

Vowels

Individual sounds that are not consonants. For examples, /a/, /e/, /i/, /o/, & /u/.

How familiar are you with English vowels?

1 2 3 4 5

1 = Not familiar at all *5 = Very familiar*

How much instruction have you received on how to produce English vowels?

1 2 3 4 5

1 = No instruction at all *5 = A lot of instruction*

When speaking English, how aware are you of how you are producing vowel sounds?

1 2 3 4 5

1 = Not aware at all *5 = Very aware*

How important are vowels in producing English speech that is understandable?

1 2 3 4 5

1 = Not important at all *5 = Very important*

Word Stress

Giving extra emphasis to a syllable in an English word. For example, "comPUter" is correct, but "COMputer" is incorrect.

How familiar are you with English word stress?

1	2	3	4	5
1 = Not familiar at all				*5 = Very familiar*

How much instruction have you received on English word stress?

1	2	3	4	5
1 = No instruction at all				*5 = A lot of instruction*

When speaking English, how aware are you of where you place stress in words?

1	2	3	4	5
1 = Not aware at all				*5 = Very aware*

How important is word stress in producing English speech that is understandable?

1	2	3	4	5
1 = Not important at all				*5 = Very important*

Intonation

The melody of English speech, or how pitch of the voice goes up and down when speaking. For example, in a yes/no question, English pitch goes up at the end of the question.

How familiar are you with English intonation?

1	2	3	4	5
1 = Not familiar at all				*5 = Very familiar*

How much instruction have you received on English intonation?

1	2	3	4	5
1 = No instruction at all				*5 = A lot of instruction*

When speaking English, how aware are you of your intonation usage?

1	2	3	4	5
1 = Not aware at all				*5 = Very aware*

How important is intonation in producing English speech that is understandable?

1	2	3	4	5
1 = Not important at all				*5 = Very important*

Rhythm

The regular beat (like in music) created by stressed elements across a sentence. These stressed elements are often content words, such as nouns and verbs. For example, "the DOG RAN to the PARK". 'Dog' (noun) 'Ran' (verb), and 'Park' (noun) are all content words, and receive extra emphasis.

How familiar are you with English speech rhythm?

1	2	3	4	5
1 = Not familiar at all				*5 = Very familiar*

How much instruction have you received on English speech rhythm?

1	2	3	4	5
1 = No instruction at all				*5 = A lot of instruction*

When speaking English, how aware are you of your speech rhythm?

1	2	3	4	5
1 = Not aware at all				*5 = Very aware*

How important is speech rhythm in producing English speech that is understandable?

1	2	3	4	5
1 = Not important at all				*5 = Very important*

Speech Rate

How slowly or quickly a person speaks English.

How aware are you of the speech rate of an English speaker you are listening to?

1	2	3	4	5
1 = Not aware at all				*5 = Very aware*

Have you ever received instruction on English speech rate?

1	2	3	4	5
1 = No instruction at all				*5 = A lot of instruction*

When speaking English, how aware are you of your speech rate?

1	2	3	4	5
1 = Not aware at all				*5 = Very aware*

How important is speech rate in producing English speech that is understandable?

1	2	3	4	5
1 = Not important at all				*5 = Very important*

APPENDIX F

Questionnaire A – Speaker Background

Name: _____ Age: _____
Hometown: _____
(City, Province/State, Country)

1) What is your native language? _____
What was your mother's native language? _____
What was your father's native language? _____
Did you speak any other languages at home? _____

2) What do you consider your second language? _____
Do you speak any other languages? _____

3) Places You Have ***Lived*** (not visited)

Location (City, Province, Country)	Reason	Length

MSU Study Experience

4) What is your intended/acquired major? _____
5) What type of degree is this (ex., BA, BSc, MA, PhD)? _____
6) How many years have you/did you studied? _____
7) Do you have a scholarship/fellowship? Yes No
If yes, could you please give a description? _____
_____
_____

Language Use and Background

8) What age did you begin to learn English? _____
9) Have you ever studied English abroad? Yes No
Where (how long)? _____
_____

10) Using the below scale, please rate your ability to speak, listen, read, and write English.

(1 = Low Ability, 9 = High Ability)

Speaking	1	2	3	4	5	6	7	8	9
Listening	1	2	3	4	5	6	7	8	9
Reading	1	2	3	4	5	6	7	8	9
Writing	1	2	3	4	5	6	7	8	9

11) If you remember, what was your last score on a test like IELTS, TOEFL, or TOEIC?
Which test? _____ Score? _____
12) If you remember, what was your score on the MSUELT? _____
13) Here in Michigan, approximately what percent of the time do you speak English (as opposed to other languages) in your daily life?

0% 10 20 30 40 50 60 70 80 90 100%

14) Here in Michigan, approximately what percent of the time do you listen to the English language media (as opposed to the media in other languages)?

0% 10 20 30 40 50 60 70 80 90 100%

15) Of the time that you spend speaking English in Michigan, approximately what percent of the time do you interact with native English speakers (as opposed to non-native speakers)?

0% 10 20 30 40 50 60 70 80 90 100%

16) Of the time that you spend speaking English in Michigan, approximately what percent of the time do you interact with nonnative English speakers (as opposed to native speakers)?

0% 10 20 30 40 50 60 70 80 90 100%

17) Please list which types of accented-English (native and nonnative) you are most familiar with.

_____

Appendix G

Questionnaire B – Listener Background

Pre-Listening

1) Please type your name.
2) Please type your e-mail address.
3) What [TESOL minor] courses are you enrolled in?
4) How old are you (in years)?
5) In what country did you study for:
    a) Elementary school?
    b) Junior high school?
    c) Senior high school?
    d) Undergraduate studies?
6) What is your current degree?
    a) Undergraduate
    b) Graduate (MA)
    c) Graduate (PhD)
    d) Other
        a. If other, please describe.
7) What is your first language? If you grew up bilingual, please list both languages.
8) What is your second language? Write none if you do not speak a second language.
9) Please rate your proficiency in your L2. (1 = Near beginner, 9 = Near nativelike)
10) Please list all other languages you speak. Please rate your proficiency on a 9-point scale (1 = Near beginner, 9 = Near nativelike).
11) Do you have previous experience <u>teaching</u> a second language? Yes or No
    a) If yes, which language(s) did you teach?
    b) How long did you teach for (in months/years)? Please respond for each language listed above.
    c) How old were your learners? Please respond for each language listed above.
    d) In what country did you teach? Please respond for each language listed above.
12) How familiar are you with the following English accents? (1 = Not familiar at all, 9 = Very familiar)
    a) American
    b) Arabic
    c) Australian
    d) British
    e) Chinese
    f) French
    g) Hindi
    h) Japanese
    i) Korean
    j) Spanish
    k) Vietnamese

Post-Listening

1) Have you ever studied the Chinese language? If you are a native speaker of Chinese, please select "No".
   a) If yes, please self-rate your proficiency below (1 = low ability, 9 = high ability).
      a. Speaking
      b. Listening
      c. Reading
      d. Writing
   b) If yes, in 2-3 sentences, please describe your Chinese learning experience (e.g., years of study, class type, location, etc.).

2) Have you ever studied the Japanese language? If you are a native speaker of Japanese, please select "No".
   a) If yes, please self-rate your proficiency below (1 = low ability, 9 = high ability).
      e. Speaking
      f. Listening
      g. Reading
      h. Writing
   b) If yes, in 2-3 sentences, please describe your Japanese learning experience (e.g., years of study, class type, location, etc.).

3) What is your major?

4) If you have a minor, please list it here?

5) Please list any linguistic courses you have taken. Please include a descriptive name, such as syntax, morphology, phonology, etc.

Appendix H

Questionnaire C – Rater Background

Name: _____ Age: _____
Hometown: _____
(City, Province/State, Country)

1) What is your native language? _____
What was your mother's native language? _____
What was your father's native language? _____
Did you speak any other languages at home? _____
2) What do you consider your second language? _____
Do you speak any other languages? _____

3) Places You Have ***Lived*** (not visited)

Location (City, Province, Country)	Reason	Length

MSU Study Experience
4) What is your current degree and focus of study? _____
- What year of study are you in? _____

5) Do you have a scholarship/fellowship? Yes No
- If yes, could you please give a description? _____

_____
6) What is your highest degree earned and focus of study? _____
- Where did you earn this degree? _____

7) Please list any other degrees earned, the focus of the degree, and location of degree. _____
_____

Language Use and Background
8) What age did you begin to learn your second language (as described above)? _____
9) Have you ever studied this language abroad? Yes No
Where (how long)? _____
_____

10) Using the below scale, please rate your ability to speak, listen, read, and write in your L2.

(1 = Low Ability, 9 = High Ability)

Speaking	1	2	3	4	5	6	7	8	9
Listening	1	2	3	4	5	6	7	8	9
Reading	1	2	3	4	5	6	7	8	9
Writing	1	2	3	4	5	6	7	8	9

11) If you remember, what was your last score on a language proficiency test (such as ACTFL)?
Which test? _____ Score? _____

12) Here in Michigan, approximately what percent of the time do you speak this language (as opposed to other languages) in your daily life?

0% 10 20 30 40 50 60 70 80 90 100%

13) Here in Michigan, approximately what percent of the time do you listen to media in this language (as opposed to the media in other languages)?

0% 10 20 30 40 50 60 70 80 90 100%

14) Of the time that you spend speaking this language in Michigan, approximately what percent of the time do you interact with native speakers (as opposed to non-native speakers)?

0% 10 20 30 40 50 60 70 80 90 100%

15) Of the time that you spend speaking this language in Michigan, approximately what percent of the time do you interact with nonnative speakers (as opposed to native speakers)?

0% 10 20 30 40 50 60 70 80 90 100%

Language Teaching Background

16) Do you have previous experience <u>teaching</u> a second language? Yes or No

l) If yes, which language(s) did you teach? _____

m) How long did you teach for (in months/years)? Please respond for each language listed above. _____

n) How old were your learners? Please respond for each language listed above. _____

o) In what country did you teach? Please respond for each language listed above. _____

Nonnative-English Speech Familiarity

17) How familiar are you with the following English accents? (1 = Not familiar at all, 9 = Very familiar)

p) American _____
q) Arabic _____
r) Australian _____
s) British _____
t) Chinese _____
u) French _____
v) Hindi _____
w) Japanese _____
x) Korean _____
y) Spanish _____
z) Vietnamese _____

18) How familiar are you with accented-English in general? (1 = Not familiar at all, 9 = Very familiar) _____

APPENDIX I

Listeners' Self-Perception of Rating Categories

How well did you understand this rating category?

- 1 = I did not understand this concept well
- 5 = Neutral
- 9 = I understood this concept well

	1	2	3	4	5	6	7	8	9
Accentedness									
Comprehensibility									
Intelligibility									

How comfortable did you feel rating this category?

- 1 = Very difficult
- 5 = Neutral
- 9 = Very easy and comfortable

	1	2	3	4	5	6	7	8	9
Accentedness									
Comprehensibility									
Intelligibility									

APPENDIX J

Paired Assessment Rating Rubric (Reproduced as presented in Ockey, 2011)

	Pronunciation Think about: a) pronunciation, b) intonation, c) word blending	**Fluency** Think about: a) automatization, b) fillers, c) speaking speech	**Grammar** Think about: a) use of morphology, b) complexity of syntax (embedded clauses, parallel structures, connectors)	**Vocabulary** Think about: a) range of vocabulary	**Communicative skills/strategies** Think about: a) interaction, b) confidence, c) conversational awareness
4	Speaks with excellent pronunciation and intonation; has practically mastered the sound system of English	Excellent fluency; uses fillers effectively; shows ability to speak quickly in short bursts	Uses both simple and complex grammar effectively; may make occasional errors but they are only in late-acquired grammar	Shows evidence of a wide range of vocabulary knowledge	Confident and natural; asks others to expand on views; shows how own and others' ideas are related; interacts smoothly
3.5 3.0	Pronunciation is good but has still not mastered the sound system of English; accent does not interfere with comprehension; can blend words	May use some fillers; rarely gropes for words but speech may still not be quick	Shows ability to use some complex grammar; may make errors but they are only in late-acquired grammar	Shows some evidence of some advanced vocabulary	Generally confident; responds appropriately to others' opinions; shows ability to negotiate meaning quickly and relatively naturally
2.5 2.0	May not have mastered some difficult sounds of English, but would	Speech is hesitant; some groping for words and unfilled spaces are	Relies mostly on simple (but appropriate) grammar; has enough	Generally has enough lexis for expressing some	Responds to others without long pauses to maintain

	be mostly understandable to a naïve NS; makes some attempts to blend words	present but generally don't impede communication completely	morphosyntax to express meaning complex grammar is attempted but may be inaccurate	opinions but does not demonstrate any particular knowledge of vocabulary	interaction; shows agreement or disagreement with others' opinions
1.5 1.0	Somewhat non-nativelike pronunciation; does not blend words together; they are pronounced in isolation	Slow, strained speech; constant groping for words and long unnatural pauses; communication with a NS would be difficult	Doesn't have enough grammar to express an opinion clearly; makes frequent errors; no attempt at complex grammar	Lexis not adequate for task; cannot express opinion properly with limited words used	Does not initiate interaction; produces monologue only; shows some turn-taking; may say, "I agree with you," but not relate ideas in explanation; too nervous to interact effectively
0.5	Very heavy accent; uses non-nativelike phonology and rhythm; words are not blended together	Fragments of speech that are so halting that conversation is not really possible; NS would think person had virtually no English	Does not use any discernible grammatical morphology	Shows knowledge of only the simplest words and phrases taught in early language learning contexts	Shows no awareness of other speakers; may speak, but not in a conversation-like way

APPENDIX K

Targeted 11 Linguistic Measures of L2 Speech

All measures are drawn from Isaacs and Trofimovich (2012) but have been relabeled to allow for increased readability.

Phonology. A total of six categories at segmental and suprasegmental levels were used to analyze the phonological properties of each speaker's speech.

(1) *Segmental Accuracy*: The total number of segmental (vowel, consonant) substitutions divided by the total number of segments articulated (e.g., substituting /i/ for / I/ in 'big') .

(2) *Syllable Structure Accuracy*: The total number of vowel and consonant epenthesis (insertion) and elision (deletion) errors over the total number of syllables articulated.

- REMOVED FROM ANALYSIS

(3) *Word Stress Accuracy*: The total number of instances where primary stress was misplaced or missing over the total number of polysyllabic words produced. (e.g., 'ciTY rather than 'CIty').

(4) *Rhythm*: A measure of English stress-timing, the number of correctly reduced syllables in both polysyllabic words and function words divided by the total number of obligatory vowel reduction contexts.

- REMOVED FROM ANALYSIS

(5) *Intonation*: The number of correct pitch patterns at the end of phrases over the total number of instances where pitch patterns were expected (e.g., in 'a man and

a woman encounter at the corner [RISING] and hit each other [FALLING]' there is one inappropriate rise after corner and one appropriate fall after other).

Fluency. Six categories designed to describe dysfluencies in L2 speech were used to measure each speaker's fluency.

(6) *Filled Pauses*: Total number of non-lexical pauses (i.e., uh, um) longer than 400 milliseconds

(7) *Unfilled Pauses*: Total number of unfilled pauses longer than 400 milliseconds

(8) *Pause Appropriateness*: A measurement of the relationship between fluency and sentence structure, the number of inappropriately filled and unfilled pauses divided by the number of total pauses produced. (e.g., in 'A restaurant that (Unfilled Pause) I enjoyed going to is Omi sushi' there is an inappropriate pause that occurs inside the phrase 'that I enjoyed').

(9) *Repetitions/Self-Corrections*: The sum of all immediately repeated and self-corrected words over the total number of words produced (e.g., 'There are (unfilled pause) big big [REPETITION] building on the (filled pause) uh intersection').

(10) *Articulation Rate*: Excluding dysfluencies (e.g., filled pauses, false starts), the total number of syllables produced divided by the total duration of the speech sample in seconds.

(11) *Mean Length of Run*: The mean number of syllables produced between two adjacent filled or unfilled pauses greater than 400 milliseconds.

REFERENCES

REFERENCES

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning, 59*(2), 249-306.

Appiah, K. A. (2006). *Cosmopolitanism: Ethics in a world of strangers (issues of our time)*. New York: W.W. Norton & Company, Inc.

Baker, A., & Murphy, J. (2011). Knowledge base of pronunciation teaching: Staking out the territory. *TESL Canada Journal, 28*(2), 29.

Baker, W. (2015). *Culture and identity through English as a lingua franca*. Berlin, Germany: De Gruyter Mouton.

Benrabah, M. (1997). Word-stress–a source of unintelligibility in English. *International Review of Applied Linguistics in Language Teaching, 35*(3), 157-166.

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

Bergeron, A., & Trofimovich, P. (2017). Linguistic dimensions of accentedness and comprehensibility: Exploring task and listener effects in second language French. *Foreign Language Annals, 50*(3), 547-566.

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.

Bowles, M. A., Toth, P. D., & Adams, R. J. (2014). A comparison of L2-L2 and L2-heritage learner interactions in Spanish language classrooms. *Modern Language Journal, 92*(2), 497-517.

Breitkreutz, J., Derwing, T. M., & Rossiter, M. J. (2001). Pronunciation teaching practices in Canada. *TESL Canada Journal, 19*, 51-61.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*(3), 341-366.

Bueno-Alastuey, M. C. (2013). Interactional feedback in synchronous voice-based computer mediated communication: Effect of dyad. *System, 41*(3), 543-559.

Byram, M. (1997). *Teaching and assessing intercultural communicative competence*. Clevedon, UK: Multilingual Matters.

Byrnes, H. (2013). Notes from the editor. *The Modern Language Journal, 97*(4), 825-827.

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). Newbury House.

Caspers, J. (2010). The influence of erroneous stress position and segmental errors on intelligibility, comprehensibility and foreign accent in Dutch as a second language. *Linguistics in the Netherlands, 27,* 17-29.

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching pronunciation: A course book and reference guide* (2nd ed.). Cambridge: Cambridge University Press.

Chambers, F. (1997). What do we mean by fluency? *System, 25*(4), 535-544.

Crowther, D., & De Costa, P. I. (2017). Developing mutual intelligibility and conviviality in the 21st century classroom: Insights from English as a lingua franca and intercultural communication. *TESOL Quarterly, 51*(2), 450-460.

Crowther, D., Kim, S., Lee, J., Lim, J., & Loewen, S. (forthcoming). Methodological synthesis of cluster analysis in second language acquisition research.

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility: Nonnative listeners' perspectives. *Journal of Second Language Pronunciation, 2*(2), 160-182.

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015a). Does a speaking task affect second language comprehensibility? *The Modern Language Journal, 99,* 80-95.

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2017). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition*. Published online 22 August 2017.

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015b). Second language comprehensibility revisited: Investing the effects of learner background. *TESOL Quarterly, 49*(4), 814-837.

Crystal, D. (2008). Two thousand million? *English Today, 24,* 3-6.

Csepes, I. (2009). *Measuring oral proficiency through paired-task performance*. Frankfurt, Germany: Peter Lang.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research, 28*(3), 369-382.

Dauer, R. M. (2005). The lingua franca core: A new model for pronunciation instruction? *TESOL Quarterly, 39*(3), 543-550.

Davies, A. (2003). *The native speaker: Myth and reality*. Clevedon, UK: Multilingual Matters.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367-396.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning, 63*(2), 163-185.

Derwing, T. M., & Munro, M. J. (2014). Training native speakers to listen to L2 speech. In J. M. Levis & A. Moyer (Eds.), *Social dynamics in second language accent* (pp. 219-236). Boston, MA: Walter de Gruyter Inc.

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Philadelphia, PA: John Benjamins.

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics, 29*(3), 359-380.

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31*(4), 553-557.

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., Munro, M. J., & Thomson, R. I. (2004). Second language fluency judgments on different tasks. *Language Learning, 54*(4), 655-679.

Deterding, D. (2010). Norms for pronunciation in Southeast Asia. *World Englishes, 29*(3), 364-377.

Ducasse, A. and Brown, A. (2009). Assessing paired orals: Rater's orientation to interaction. *Language Testing, 26*(3), 423-443.

Dziubalska-Kołaczyk, K., & Przedlacka, J. (Eds.). (2005). *English pronunciation models: A changing scene*. New York, USA: Peter Lang.

Educational Testing Service. (2012). *The official guide to the TOEFL test* (4th ed.). New York: McGraw Hill.

Educational Testing Service. (2017). *The official guide to the TOEFL test* (5th ed.). New York: McGraw Hill.

Educational Testing Service. (2014). *TOEFL iBT speaking section scoring guide*. Retrieved online from https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf.

Everitt, B. S. (1980). *Cluster analysis* (2nd ed.). New York: Halsted Press.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*(3), 313-326.

Field, A. (2009). *Discovering statistics using SPSS* (3rd Ed.). London: SAGE Publishing.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39*(3), 399-423.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Effects of age of second-language learning on the production of English consonants. *Speech Communication, 16*, 1-26.

Foote, J. A., Holtby, A. K., & Derwing, T. M. (2011). Survey of teaching pronunciation in adult ESL programs in Canada, 2010. *TESL Canada Journal, 29*, 1-22.

Foote, J. A., Trofimovich, P., Collins, L., & Urzúa, F. S. (2016). Pronunciation teaching practices in communicative second language classes. *The Language Learning Journal, 44*(2), 181-196.

Fowler, C. A., Brown, J., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language, 49*(3), 296–314.

Fulcher, G., & Owens, N. (2016). Dealing with the demands of language testing and assessment. In G. Hall (Ed.), *The Routledge Handbook of English Language Teaching* (pp. 109-120). New York: Routledge.

Galaczi, E. D. (2008). Peer–peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5*(2), 89-119.

Galaczi, E., Post, B., Li, A., Barker, F., & Schmidt, E. (2017). Assessing second language pronunciation: Distinguishing features of rhythm in learner speech at different proficiency levels. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment*: *Interdisciplinary perspectives* (pp. 157-182). Bristol, UK: Multilingual Matters.

Gallois, C., Ogay, T., & Giles, H. (2005). Communication accommodation theory: A look back and a look ahead. In W. B. Gudykunst (Ed.), *Theorizing about intercultural communication* (pp. 121-148). Thousand Oaks, CA: SAGE Publications.

Galloway, N., & Rose, H. (2015). *Introducing global Englishes*. New York: Routledge.

Garrod, S., & Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. *Topics in Cognitive Science, 1*(2), 292-304.

Gass, S., & Mackey, A. (2015). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 180-206). New York: Routledge.

Gass, S. M., & Mackey, A. (2017). *Stimulated recall methodology in second language research* (2nd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34*, 65-89.

Ginther, A., & Elder, C. (2014). A comparative investigation into understandings and uses of the TOEFL iBT® test, the International English Language Testing Service (Academic) test, and the Pearson Test of English for graduate admissions in the United States and Australia: A case study of two university contexts. *ETS Research Report Series, 2014*(2), 1-39.

Gurzynski-Weiss, L., & Baralt, M. (2014). Exploring learner perception and use of task-based interactional feedback in FTF and CMC modes. *Studies in Second Language Acquisition, 36*, 1-37.

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing*, *29*(2), 163-180.

Harding, L. (2017). What do raters need in a pronunciation scale? The users' view. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment* (pp. 12-34). Bristol, UK: Multilingual Matters.

Harding, L. (2018). Validity in pronunciation assessment. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 30-48). New York: Routledge.

Hardison, D. (2014). Phonological literacy in L2 learning and teaching. In J. M. Levis & A. Moyer (Eds.), *Social dynamics in second language accent* (pp. 195-218). Boston, MA: Walter de Gruyter Inc.

Hardison, D. M. (in press). Visualizing the acoustic and gestural beats of emphasis in multimodal discourse: Theoretical and pedagogical implications. *Journal of Second Language Pronunciation*.

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.

Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *The Language Learning Journal, 36*(2), 153-166.

International English Language Testing System. (2009). *Cambridge IELTS 7: Examination papers from University of Cambridge ESOL Examinations: English for speakers of other languages.* Cambridge: Cambridge University Press.

International English Language Testing System. (2011). *Cambridge IELTS 8: Examination papers from University of Cambridge ESOL Examinations: English for speakers of other languages.* Cambridge: Cambridge University Press.

International English Language Testing System. (2016). *Speaking assessment criteria*. Retrieved from https://www.ielts.org/~/media/pdfs/speaking-band-descriptors.ashx.

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

Issacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition, 34*(3), 475-505.

Issacs, T., & Trofimovich, P. (Eds). (2017). *Second language pronunciation assessment.* Bristol, UK: Multilingual Matters.

Isaacs, T., Trofimovich, P., & Foote, J. A. (2017). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing.* Published online 6 May 2017.

Isaacs, T., Trofimovich, P., Yu, G., & Muńoz Chereau, B. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS Pronunciation scale. *IELTS Research Reports Series, 4*, 1-48.

Isbell, D. (2018). Assessing pronunciation for research purposes with listener-based numerical scales. In O. Kang & A. Ginther (Eds.), *Assessment of second language pronunciation* (pp. 89-112). New York: Routledge.

Isbell, D., Park, O.–S., & Lee, K. (in press). Learning Korean pronunciation: Effects of instruction, proficiency, and L1. *Journal of Second Language Pronunciation.*

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford, UK: Oxford University Press.

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

Jenkins, J. (2006). Current perspectives on teaching world Englishes and English as a lingua franca. *TESOL Quarterly, 40*, 157-181.

Jenkins, J. (2014). *English as a lingua franca in the international university*. New York: Routledge.

Jewitt, C. (2014). An introduction to multimodality. In C. Jewitt (Ed.), *The Routledge Handbook of Multimodal Analysis* (2nd Ed.) (pp. 15-30). New York, NY: Routledge.

Johnson, M., & Tyler, A. (1999). Re-analyzing the OPI: How much does it look like natural conversation? In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 27-52). Philadelphia, PA: John Benjamins North America.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning, 64*(4), 809-854.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System, 38*(2), 301-315.

Kang, O., & Ginther, A. (Eds.) (2018). *Assessment in second language pronunciation*. New York: Routledge.

Kang, O., & Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology, 28*(4), 441-456.

Kang, O., Rubin, D. L., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *Modern Language Journal, 94*(4), 554-566.

Kennedy, S., Guénette, D., Murphy, J., & Allard, S. (2015). Le rôle de la prononciation dans l'intercompréhension entre locuteurs de français lingua franca [The role of pronunciation in comprehension between speakers of French as a lingua franca]. *Canadian Modern Language Review, 71*, 1-25.

King, R. S. (2015). *Cluster analysis and data mining: An introduction*. Boston, MA: Mercury Learning and Information.

Knapp, M. L., & Hall, J. A. (1992). *Nonverbal communication in human interaction*. New York, NY: Holt Rinehart and Winston, Inc.

Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing, 16*(2), 163-188.

Kumaravadivelu, B. (2008). *Cultural globalization and language education*. New Haven, CT: Yale University Press.

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

Lazarton, A., & Davis, L. (2008). A microanalytic perspective on discourse, proficiency, and identity in paired oral assessment. *Language Assessment Quarterly, 5*(4), 313-335.

Leaper, D. A., & Riazi, M. (2014). The influence of prompts on group oral tests. *Language Testing, 31*(2), 177-204.

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics, 36*(3), 345-366.

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

Leung, C. (2005). Convivial communication: Reconceptualizing communicative competence. *International Journal of Applied Linguistics, 15*(2), 119-144.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369-377.

Loewen, S. (2015). *Introduction to instructed second language acquisition*. New York: Routledge.

Loewen, S., & Isbell, D. (2017). Pronunciation in face-to-face and audio-only synchronous computer-mediated learner interactions. *Studies in Second Language Acquisition, 39*(2), 225-256.

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

Long, M. H., & Porter, P. (1985). Group work, interlanguage talk, and second language acquisition. *TESOL Quarterly, 19*(2), 207-228.

Low, E.–L. (2015). *Pronunciation for English as an international language: From research to practice*. New York: Routledge.

Mackey, A., Gass, S., & McDonough, K. (2000). How do learners' perceive interactional feedback? *Studies in Second Language Acquisition, 22*(4), 471-497.

Mackey, A. & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407-452). Oxford: Oxford University Press.

MacKay, I. R. A., Flege, J. E., & Imai, S. (2006). Evaluating the effects of chronological age and sentence duration on degree of perceived foreign accent. *Applied Psycholinguistics, 27*(2), 157–183.

MacKenzie, I. (2011). *Intercultural negotiations*. New York: Routledge.

Major, R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology*. Mahwah, NJ: Lawrence Erlbaum.

Major, R. C., Fitzmaurice, S. F., Bunta, F., Balasubramanian, C. (2001). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly, 36*(2), 173-190.

Marsden, E., Mackey A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS Repository of Instruments for Research into Second Languages* (pp. 1-21). New York: Routledge.

Matsuda, A. (Ed.). (2017). *Preparing teachers to teach English as an international language*. Bristol, UK: Multilingual Matters.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*(3), 397-421.

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2), 127-145.

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

Moyer, A. (2013). *Foreign accent: The phenomenon of non-native speech.* Cambridge: Cambridge University Press.

Munro, M. J. (2018, forthcoming). Dimensions of pronunciation. In O. Kang, R. I. Thomson, & J. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation*. London: Routledge.

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

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49*(s1), 285-310.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition, 23*(4), 451-468.

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

Munro, M. J., Derwing, T. M., & Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication, 52*(5-7), 626-637.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28*, 111-131.

Nagle, C. (2018). Motivation, comprehensibility, and accentedness in L2 Spanish: Investigating motivation as a time-varying predictor of pronunciation development. *Modern Language Journal, 102*, 199-217.

Nakatsuhara, F. (2006). The impact of proficiency-level on conversational styles in paired speaking tests. *Cambridge ESOL Research Notes, 25*, 15-20.

Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing, 31*(2), 147-175.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language learning, 50*(3), 417-528.

Norton, B. (2013). *Identity and language learning* (2nd ed.). Toronto, Canada: Multilingual Matters.

Norton, J. (2005). The paired format in the Cambridge Speaking Tests. *ELT Journal, 59*(4), 287–297.

O'Brien, M. G. (2014). Learners' assessment of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning, 64*(4), 715-748.

O'Brien, M. G. (2016). Methodological choices in rating speech samples. *Studies in Second Language Acquisition, 38*(3), 587-605.

Ockey, G. J. (2009). The effects of group members/ personalities on a test taker's L2 group oral discussion test scores. *Language Testing, 26*(2), 161-186.

Ockey, G. (2011). Self-consciousness and assertiveness as explanatory variables in of L2 oral ability: A latent variable approach. *Language Learning, 61*(3), 968-989.

Ockey, G. J., & French, R. (2014). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics, 37*(5), 693-715.

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing, 32*, 39-62.

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Science, 12*(6), 237-241.

Park, J. S-Y. & Wee, L. (2015). English as a lingua franca: Lessons for language and mobility. In C. Stroud & M. Prinsloo (Eds.), *Language, literacy and diversity: Moving words* (pp. 55-71). New York: Routledge.

Pickering, L. (2009). Intonation as a pragmatic resource in ELF interaction. *Intercultural Pragmatics, 6*(2), 235–255.

Pickering, L., & Litzenberg, J. (2011). Intonation as a pragmatic resource in ELF interaction, revisited. In A. Archibald, A. Cogo, & J. Jenkins (Eds.), *Latest trends in ELF research* (pp. 77–92). Newcastle, UK: Cambridge Scholars Publishing.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition, 35*(4), 655-687.

Plonsky, L. (Ed.). (2015a). *Advancing quantitative methods in second language research*. New York: Routledge.

Plonsky, L. (2015b). Statistical power, *p* values, descriptive statistics, and effect sizes: A "back-to-basics" approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 23-45). New York: Routledge.

Plonsky, L., & Derrick, D. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal, 100*(2), 538-553.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning, 61*(2), 325-366.

Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning, 65*(s1), 9-36.

Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (forthcoming). A methodological synthesis of judgment tasks in second language research.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912.

Plough, I. C., & Bogart, P. S. (2008). Perceptions of examiner behavior modulate power relations in oral performance testing. *Language Assessment Quarterly, 5*(3), 195-217.

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

Robinson, P. (2011). Task-based language learning: A review of issues. *Language Learning, 61*, 1–36.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduate's judgments of nonnative English-speaking teaching assistants. *Research in Higher Education, 33*(4), 511-531.

Saito, K. (2012). Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies. *TESOL Quarterly, 46*(4), 842-854.

Saito, K., & Akiyama, Y. (2017). Linguistic correlates of comprehensibility in second language Japanese speech. *Journal of Second Language Pronunciation, 3*(2), 199-217.

Saito, Y., & Saito, K. (2017). Differential effects of instruction on the development of second language comprehensibility, word stress, rhythm, and intonation: The case of inexperienced Japanese EFL learners. *Language Teaching Research, 21*(5), 589-608.

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics 37*(2), 217-240.

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics, 38*(4), 439-462.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness, and sense relations. *Studies in Second Language Acquisition, 38*(4), 677-701.

Sato, M. (2014). Exploring the construct of interactional oral fluency: Second language acquisition and language testing approaches. *System, 45*, 79-91.

Scollon, R., Scollon, S. W., & Jones, R. H. (2012). *Intercultural communication: A discourse approach* (3rd ed.). Oxford: Blackwell.

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

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching, 54*(2), 79-95.

Seidlhofer, B. (2011). *Understanding English as a lingua franca*. Oxford: Oxford University Press.

Sewell, A. (2017). Functional load revisited: Reinterpreting the findings of 'lingua franca' intelligibility studies. *Journal of Second Language Pronunciation, 3*, 57-79.

Sifakis, N.C., & Sougari, A.-M. (2005). Pronunciation issues and EIL pedagogy in the periphery: A survey of Greek state school teachers' beliefs. *TESOL Quarterly, 39*(3), 467-488.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics, 30*(4), 510–32.

Smith, B. L., & Hayes-Harb, R. (2011). Individual differences in the perception of final consonant voicing among native and non-native speakers of English. *Journal of Phonetics, 39*, 115-120.

Solon, M., Long, A. Y., & Gurzynski-Weiss, L. (2017). Task complexity, language-related episodes, and production of L2 Spanish vowels. *Studies in Second Language Acquisition, 39*(2), 347-380.

Southwood, M. H., & Flege, J. E. (1999). Scaling foreign accent: Direct magnitude estimation versus interval scaling. *Linguistics & Phonetics, 13*(5), 335-349.

Staples, S., & Biber, D. (2015). Cluster analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 243-274). New York: Routledge.

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning, 52*, 119–58.

Sueyoshi, A., & Hardison, D. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*(4), 661-669.

Suvorov, R. (2011). *The effects on context visuals on L2 listening comprehension*. Cambridge ESOL: Research Notes, 45, 2-7.

Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing, 32*(4), 463-483.

Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal, 82*(3), 320-337.

Thompson, G. L., Cox, T. L., & Knapp, N. (2016). Comparing the OPI and the OPIc: The effect of test method on oral proficiency scores and student preference. *Foreign Language Annals, 49*, 75-92.

Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning, 41*(2), 177-204.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effects of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*, 1-30.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*(4), 905-916.

VanPatten, B., & Williams, J. (2015). *Theories in second language acquisition: An introduction*. New York: Routledge.

Varonis, E. M., & Gass, S. M. (1982). The comprehensibility of nonnative speech. *Studies in Second Language Acquisition, 4*(2), 114-136.

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning and Technology, 11*, 67–86.

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, *5*(3), 218-243.

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

Warner, R. M. (2008). *Applied statistics: From bivariate through multivariate techniques*. London: SAGE Publications.

Winke, P. (2014). Testing hypotheses about language learning using structural equation modeling. *Annual Review of Applied Linguistics, 34*, 102-122.

Winke, P., & Gass, S. (2013). The influence of L2 experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly*, *47*(4), 762-789.

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.

Yan, X., & Ginther, A. (2018). Listeners and raters: Similarities and differences in evaluation of accented speech. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 67-88). New York: Routledge.

Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkle (Ed.), *Handbook of research in second language learning* (Vol. 2, pp. 426-443). New York, NY: Routledge.

Yule, G., & Macdonald, D. (1990). Resolving referential conflicts in L2 interaction: The effect of proficiency and interactive role. *Language Learning, 40*(4), 539-556.