Towards Defining a Valid Assessment Criterion of Pronunciation Proficiency in Non-Native English Speaking Graduate Students

Talia Isaacs, Department of Integrated Studies in Education

McGill University, Montreal

A thesis submitted to McGill University in partial fulfillment
of the requirements of the degree of Master of Arts

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-24877-5
Our file Notre référence
ISBN: 978-0-494-24877-5

Acknowledgements

This thesis could not have been completed without the help and support of several individuals. Dr. Carolyn Turner, my thesis supervisor, personifies resourcefulness and is a true mentor. Dr. Eddy and Honey Isaacs were gracious enough to help me revamp my rating scale from seven time zones away. Dr. Pavel Trofimovich initiated me into the vast and opaque world of sound editing one sunny afternoon. Dr. Wayne Dickerson, whose definition of intelligibility is central to this thesis, sent me a most thorough response to my questions via e-mail. Information furnished by Dr. Sarah Briggs following her *LTRC* talk lent insight into language testing at an institution where pronunciation really "matters." The invaluable advice of Carolyn Samuel and Sarita Kennedy remind me that pronunciation work in Montreal need not be done in isolation. Dr. Roy Lyster, Jesús Izquierdo, and David Syncox provided me with a viable "Plan B" for my sound recordings. Randall Halter helped me sleep better at night by backing up all of my mega-sized speech files. The Educational Testing Service gave me permission to use a 1995 version of their copyrighted *Test of Spoken English®*. Jason Blank vividly depicted for me the reaction of a typical science undergraduate student to my unrevised rating scheme a few days before Dr. Joe Schwarcz gave me access to such a population. Jérémie Mbairamadji helped me understand the ABCs of data entry in SPSS. By masterminding how I could reduce the number of variables, Dr. Yanhong Zhang lessened my right arrow scrolling time and prevented undue neck strain. Lynn Gregg helped with stylistic considerations and editing strategies. Youn-Hee Kim, always one step ahead, gave me a two month glimpse into my "thesis future." Rachelle Beaudin showed me the true meaning of strength and perseverance in the face of adversity. My participants, who inspired me with their enthusiasm, have helped this study to come alive. Marcin Cudnoch has shared most closely with me this epoch of my life.

Abstract

This exploratory, mixed-design study investigates whether intelligibility is "enough," that is, a suitable goal and an adequate assessment criterion, for evaluating proficiency in the pronunciation of non-native English speaking graduate students in the academic domain. The study also seeks to identify those pronunciation features which are most crucial for intelligible speech.

Speech samples of 19 non-native English speaking graduate students in the Faculty of Education at McGill University were elicited using the *Test of Spoken English* (*TSE*), a standardized test of spoken proficiency which is often used by institutions of higher learning to screen international teaching assistants (ITAs). Results of a fined-grained phonological analysis of the speech samples coupled with intelligibility ratings of 18 undergraduate science students suggest that intelligibility, though an adequate assessment criterion, is a necessary but not a sufficient condition for graduate students to instruct undergraduate courses as teaching assistants, and that there is a threshold level (i.e., minimum acceptable level) of intelligibility that needs to be identified more precisely. While insights about the features of pronunciation that are most critical for intelligibility are inconclusive, it is clear that intelligibility can be compromised for different reasons and is often the result of a combination of "problem areas" that interact together.

The study has some important implications for ITA training and assessment, for the design of graduate student pronunciation courses, and for future intelligibility research. It also presents a first step in validating theoretical intelligibility models which lack empirical backing (e.g., Morley, 1994).

Résumé

La présente étude mixte et exploratoire cherche à savoir si l'intelligibilité à elle seule constitue un but suffisant et si elle peut servir de critère d'évaluation adéquat en matière de prononciation pour des étudiants du 2e cycle universitaire en anglais langue seconde. Cette étude cherche aussi à identifier les caractéristiques de prononciation nécessaires à l'intelligibilité du langage.

19 étudiants du 2e cycle universitaire de la faculté d'éducation de l'Université McGill, dont l'anglais est une langue seconde, ont passé un examen verbal, le « Test of Spoken English », permettant la cueillette d'échantillons linguistiques. Cet examen standardisé est communément utilisé par des écoles d'enseignement supérieur dans le choix des assistants internationaux à l'enseignement (AIE). Les résultats d'une analyse phonologique très pointue des échantillons linguistiques ainsi qu'une compilation d'indices d'intelligibilité provenant de 18 étudiants du 1er cycle universitaire en sciences révèlent que bien que l'intelligibilité soit un critère d'évaluation nécessaire, elle ne suffit pas aux besoins des étudiants du 2e cycle qui enseigne à titre de AIE au 1er cycle. En effet, il existe un seuil d'intelligibilité (c.à-d. niveau minimum acceptable) qui doit être défini de façon plus précise. Bien que notre aperçu des caractéristiques de prononciation nécessaire à l'intelligibilité demeure incomplet, nul ne peut douter du fait que plusieurs facteurs y contribuent et que souvent ceux-ci agissent ensemble pour devenir problématique.

Cette étude comporte d'importantes conséquences en matière d'évaluation et de formation des AIE, de la planification des cours de prononciation des étudiants du 2e cycle en matière de recherche future liée à l'intelligibilité. Par ailleurs, cette étude fait un premier pas vers la validation théorique de modèles d'intelligibilité qui jusqu'ici était sans appui empirique (par ex. Morley, 1994).

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

After a period of relative neglect, pronunciation teaching is making a comeback (Morley, 1991). This is evidenced by the proliferation of accent-reduction courses in North America over the last few years, a trend that is likely to continue against the backdrop of globalization and preponderance of English as the international *lingua franca*. Pronunciation teaching taps into the market niche of groups of non-native speakers whose tasks in their professional domain mandate that their speech be easily understood (e.g., lawyers, university professors, doctors). It is also fueled by a growing body of non-native speakers who have chosen to pursue higher education in English speaking countries, and who may need a sufficient command of English in order to carry out the tasks that are demanded of them in an increasingly competitive academic environment.

In pronunciation teaching and testing, the traditional focus on "accuracy" and goal of attaining native-like pronunciation has been discarded as inappropriate and unrealistic for second language learners. Evidence for this is drawn from the so-called Joseph Conrad phenomenon, the idea that it is almost impossible for adult non-native speakers to eradicate traces of foreign accent in their speech (Celce-Murcia et al., 1996). In the recent revival of pronunciation, the traditional measure of "accuracy" against the native-speaker norm has thus been replaced by the broader goal of "intelligibility."

Having entered the English lexicon from the Latin *intelligere* in the late 14th century and evolving to denote "capable of being understood" by the early 17th century, the word "intelligible" and its noun form "intelligibility" are still very contemporary (Harper, 2001).[1] Indeed, a simple *Google* search for "intelligibility," which yields over 1,340,000 hits, unearths websites for subjects as varied as church sounds and acoustics, digital cartography, plastic and reconstructive surgery, philosophical treatises on the "intelligibility of the universe," tracheotomies, and finally "keeping up with the joneses – clichés on a topic."[2]

In the field of applied linguistics, the word "intelligibility," having infiltrated the literature in second language pronunciation, world Englishes, and second language

[1] In fact, "intelligible" and "intelligent" – faculty of understanding – both derive from the same root.
[2] The search was conducted on September 7, 2005.

assessment, is also used in many different ways (perhaps even becoming a cliché).[3] Indeed, in light of the emergence of English as *the* undisputed global language (see Crystal, 2003) and the existence of many varieties of English on the international stage, Jenkins argues for the need to establish a "pronunciation core of intelligibility" as the baseline for the future (2000, p. 21). "Intelligibility" is also cited in the speaking rubrics of "the next generation *TOEFL* test" (Educational Testing Service, 2005) which, with its Star Trekesque marketing ploy,[4] ensures a bright future for intelligibility as a criterion for spoken assessment in the world of high stakes testing. Others, such as Taylor (1991), assert that English is now a means of global communication and that "the tacit assumption (in pronunciation teaching) has always been that we should aim to make learners 'intelligible'" (p. 425). The fact that discussions on "intelligibility" can be found as early as Abercrombie (1956) and Lado (1961), that is, well before the notion of English as a global language became popularized,[5] lends credence to the belief that "the fundamental linguistic virtues—simplicity, clarity, intelligibility—are unassailable, but they must be constantly reinterpreted against an evolving social and linguistic background" (Nunberg, 2000). It seems, then, that uses of the word "intelligibility" can be traced into the past, exist in the present, and are likely to continue with full force into the future.

Yet despite its widespread use, intelligibility has been differentially defined in the applied linguistics literature, and the notion of what exactly constitutes intelligibility is as yet unclear (Field, 2005). In fact, the only thing that researchers seem to agree upon is that the term is shrouded in obscurity and that a universal definition is lacking. To list a few examples, what some researchers call "intelligibility" is for others "comprehensibility," the notion of "irritability" is inherent in some people's conceptions of the term but not others, and the degree of onus that is placed on either the speaker or the listener to be intelligible (or to attain a certain level of intelligibility) is not constant across studies. (See discussions in Jenkins, 2000; Derwing & Munro, 1997).

[3] There is often considerable overlap within these subdisciplines, if we can consider them as such.
[4] Notably, the catch phrase "next generation" is opted for rather than "new generation."
[5] The debate surrounding "English as a global language," which had been festering through the late 1980's and early 90's, came to a head with the publication of Crystal's book by that title in 1997.

Comparing studies on intelligibility by different authors, then, is essentially wading in murky waters. But what *is* clear is that intelligibility is a complex construct that needs to be adequately defined for the purposes for which it will be used so as not to contribute to the confusion.

The purpose of this thesis is, broadly speaking, to shed some light on, or least turn the spotlight on intelligibility as it relates to the assessment of second language pronunciation. Before embarking on the details of the study, however, it must be situated in the body of existing literature that is not merely part of the backdrop of the study, but rather the stream of inspiration from which the inquiry springs and essential to an understanding of the research results and implications. Chapter 2 opens with a synthesis of the differential theoretical definitions of "intelligibility" in the literature and its operationalization in empirical studies. It then shifts to a discussion on the role of intelligibility in different forms of spoken assessment that are currently in use, particularly for non-native English speaking graduate students and international teaching assistants (ITAs) in the academic domain. Chapter 3 delineates the rationale and research questions of the study, the method that was employed to address the research questions, and the context in which the study was deployed. It provides the link between the theory and research questions outlined in Chapter 2 and the actual operationalization given the practical constraints of the study. A description of the research participants, the instruments which were used, and the data collection and analysis procedures are all an integral part of the chapter. This leads into Chapter 4, which presents the results of quantitative and qualitative analyses in direct response to the research questions that were posed in Chapter 3. Where possible, illustrative examples are also drawn from the phonological data, which were transcribed (segmentally and suprasegmentally) and color-coded for intelligibility. The analyses are sometimes discussed in isolation and sometimes merged to give a more holistic picture. Finally, Chapter 5 discusses strengths and weaknesses of the study and possible directions for future research.

Chapter 2: Literature Review

Towards Making "Intelligibility" More Intelligible

In an introduction to a 1989 article, Brown wrote, "intelligibility is a concept which has been widely appealed to by linguists. However, as a technical term, it does not have a precise definition subscribed to by all linguists." Echoing this over a decade later under the subject heading "what do we mean by intelligibility," Jenkins explained, "there is as yet no broad agreement on a definition of the term 'intelligibility': it can mean different things to different people" (2000, p. 69). Lado's comment some 40 years earlier, therefore, that that the criterion of intelligibility as a standard of pronunciation is difficult to define (1961) seems to be as applicable now as then.

The purpose of this review of the literature is to synthesize the different ways in which "intelligibility" has been defined and measured in the research literature insofar as it relates to second language pronunciation. A consolidation of the various interpretations of intelligibility may help to tease apart some important concepts in the field of second language pronunciation that have often been confused as a result of differences in nomenclature or in the degree of inclusiveness of the definition, on the path towards one day coming to a field-wide consensus on a definition of intelligibility in pronunciation. At the very least, this chapter sets out to unveil the various reasons as to why the construct of intelligibility has been problematized, in order better to understand the impetus behind the present study, its innovations, and its limitations.

Defining Intelligibility

The advent of the Communicative Approach in the 1980s saw a shift in focus in pronunciation teaching from "getting the message across" to "getting the sounds correct" (Yule, 1990, p. 107; Celce-Murcia et al., 1996). Abercrombie (1956) foreshadowed this paradigm shift by throwing into question whether language learners really need to acquire perfect (i.e., native-like) pronunciation. In fact, most language learners "need no more than a comfortably intelligible pronunciation" (p. 36). Abercrombie proceeded to define "'comfortably' intelligible" as "a pronunciation which can be understood with little or no conscious effort on the part of the listener." Whether or not he was cognizant at the time that his catch phrase would continue to resonate in pronunciation circuits up until the present time is uncertain.

Kenworthy (1987) shares Abercrombie's view that being "comfortably intelligible" is a "far more reasonable goal" for the majority of language learners than striving for native-like pronunciation (p. 3).[6] Indeed, latent in her conception of "comfortable intelligibility" is the broader context of communication - the fact that second language speakers "need to be intelligible so that they can communicate" (p. 15).

In order to understand "comfortable intelligibility," which Kenworthy claims is the goal of pronunciation, it is useful to pull these words apart and work with them separately. In providing "one definition" of intelligibility (as though to imply that there are many), Kenworthy articulates it as "being understood by a listener at a given time in a given situation" and equates it with "understandability" (p. 13). In search of a more operational definition, she elaborates that if the second language speaker substitutes a certain sound or feature of pronunciation for another and the listener hears a different word or phrase than the speaker had intended to say, the result is unintelligibility. Conversely, if the word is understood, then it is said to be intelligible. It follows that the more words the listener is able to accurately identify, the more intelligible the speaker is.

There are a few things to note in this interpretation of intelligibility. Kenworthy's operational definition pertains to intelligibility at the word level, where the speaker's intent in uttering a word must be contended with, although this is often hard to gauge from the listener's point of view. Notably, the first definition she provides is written in the passive voice, with a greater emphasis on the listener, the agent, than on the speaker, the unmentioned subject. Although Kenworthy makes the claim that "intelligibility has as much to do with the listener as with the speaker" (p. 14), her notion of "comfortably" also focuses mostly on the listener. If the speaker pronounces such that the listener constantly needs to ask for repetition and/ or clarification - that is, if the act of listening to a non-native speaker becomes too laborious - then the listener, having reached his/ her tolerance threshold, becomes frustrated or irritated. Being comfortably intelligible has to do with efficiency, then, where the listener can understand the speaker without too much difficulty or recourse to repetition.

[6] Although the phrase "comfortably intelligible" is not used by all pronunciation proponents, the view that intelligibility, (however it is defined), is a more readily attainable and/ or desirable goal for non-native speakers than native-like pronunciation is echoed time and time again in the pronunciation literature. (See, for example, Wong, 1987).

Morley (1994) also makes use of the term "comfortably intelligible." While she does not define the term, she does contend that unless non-native speakers are comfortably intelligible, they often avoid spoken interaction. This implies that comfortable intelligibility is something that a speaker either has or does not have. Morley continues that "speakers with poor intelligibility have long-range difficulties in developing into confident and effective oral communicators; some never do" (p. 67). This brings to mind Morley's (1991) state-of-the-art paper, in which she argued that "it is imperative that students' personal/social language needs, *including reasonably intelligible pronunciation*, be served with instruction that will give them communicative empowerment" [original emphasis] (p. 489), noting that poor pronunciation can be both professionally and socially disadvantageous for an individual.

To an even larger extent than Kenworthy (1987), then, Morley (1991; 1994) makes explicit the link between intelligibility and communication. Indeed, replete with a plethora of ideas about "new wave pronunciation" (p.70), her more recent article concludes that pronunciation be rewritten into language instruction with a "new look following the premise that *intelligible pronunciation and global communication are essential components of communicative competence*" (p. 90).[7] This seems to be a restatement of the "basic premise" in her earlier article that "*intelligible pronunciation is an essential component of communicative competence*" (p. 488), with the mere addition of the words "global communication" in the more recent quote.

Unfortunately, Morley throws around many terms and concepts in her 1994 article, and the link between the various frameworks that she presents is not always clear. In addition to "comfortable intelligibility," she also alludes to "functional intelligibility," "overall intelligibility," and just plain "intelligibility" (i.e., without any qualifier). As well, she often refers to "intelligibility" in conjunction with "communicability." It will be necessary to untangle these terms as much as possible in order to make sense of her theoretical models and get to the bottom of her conception of intelligibility.

"Functional intelligibility" is listed along with "functional communicability" as two out of four "learner speech-pronunciation goals" which are key for "achieving

[7] Both of Morley's articles allude to Canale and Swain's (1980) framework of "communicative competence." This framework is essential to her argument for the need to reintegrate pronunciation into the communicative language curriculum.

satisfactory communicative speech-pronunciation patterns" (pp. 78-79). The intent of "functional intelligibility… is to help learners develop spoken English that is (at least) reasonably easy to understand and not distracting to listeners," while "functional communicability" seeks to "help the learner develop spoken English that serves his or her individual communicative needs effectively for a feeling of communicative competence" (p. 78). Morley's notion of "functional intelligibility," then, is similar to Kenworthy's (1987) "comfortably intelligible" in terms of the ease of understanding of a non-native speaker's speech, although speech that is too difficult to understand will apparently be "irritating" for Kenworthy's listener and "distracting" for Morley's listener (1994). Also, Morley's notion of "functional intelligibility," which is part of a compendium of ideas in a loose model for pedagogues to consult (Table 4 in the article) engenders the notion of "helping the learner" whereas this feature is absent from Kenworthy's operational definition.

In a table outlining the dual focus of speech production and performance, which looks like an intense brainstorm of "micro level speech pronunciation: discrete points" in one column and "macro level features of speech performance: global patterns" in the other, Morley makes no apparent link between the items in the two columns other than that they are squished into the same table and have the same formatting (1994, p. 75). "Overall speech intelligibility" is cited as one of the seven macro level features that are listed, and the reader is directed to consult the *Speech intelligibility/ communicability index* (p. 76), a six-level framework which inextricably links intelligibility with its impact on communication. This table is conceived as an assessment tool to evaluate what Morley calls "overall intelligibility."[8]

In the first column of the index, Morley describes speech in terms of intelligibility; in the second column, she evaluates its impact on communication. Here, the link between the two columns is evident. Scale descriptors 1-5 describe speech as being "basically unintelligible," "largely unintelligible," "reasonably intelligible," "largely intelligible," and "fully intelligible" respectively, that is on an intelligibility continuum with 5 being the most intelligible and 1 the least intelligible (p. 76). Morley's

[8] Morley's (1994) *Speech intelligibility/ communicability index* is also reproduced in the appendix of Celce-Murcia et al. (1996), which is testament to its (perceived) usefulness.

notion of "functional intelligibility" underscores the descriptions that follow this "adverb + intelligible" formula that describes degree of intelligibility. While in scale descriptor 1 only occasionally can a speaker's word or phrase be recognized, in scale descriptors 2-5 listener effort and features which distract the listener are directly alluded to and placed on the continuum, with scale descriptor 2 entailing the most listener effort/ distraction and scale descriptor 5 entailing the least listener effort/ distraction.

The corresponding "impact on communication" at each scale band is described in the adjoining column in terms of the degree of interference of accent in getting the message across, with scale descriptor 1 depicting the most interference, where accent impedes functional communication, and scale descriptor 5 the least interference, in which accent does not affect speech functionality. Finally, at scale band 6, speech on the intelligibility column is described as native-like, with only minimal divergence from the native speaker norm. Presumably, no direct reference to intelligibility is made here since full intelligibility was already attained at intelligibility descriptor 5. In the impact on communication descriptor in the neighboring column, accent is correspondingly described as "virtually nonexistent" (p. 77).

To put the above in perspective, the ideas presented in Morley's chapter are brilliant in terms of their innovation and inspiring in the manner of the most rousing keynote address. Though replete with information, this "multidimensional" chapter is an excellent resource for ESL researchers and pedagogues alike, and presents an abundance of forward-thinking ideas and theoretical models from which people in the field can build. This is a crucial first step in "write(ing) pronunciation back into the instructional equation" on a large scale (Morley, 1991, p. 488), in giving pronunciation the credibility and accessibility that it requires in the worlds of pronunciation research, pedagogy, and assessment, and in inspiring aspiring pronunciation proponents to do more work in this area where it is so badly needed. It may well be that pronunciation will never again become the "Cinderella of language teaching" (Kelly, 1969, p. 87), but it also needn't any longer be the neglected "orphan in English programs around the world" (Gilbert, 1994, p. 38).

This being said, Morley's theoretical models lack empirical backing. Morley's *Speech intelligibility/ communicability index* is a prime example of this. Indeed, without any apparent justification in the chapter, Morley assigns to her index two different threshold levels. Communicative Threshold A appears before scale band 3 and Communicative Threshold B appears before scale band 5. Scale bands 1 and 2, then, can be considered pre-communicative threshold levels.

The idea of a threshold level of intelligibility is not new in pronunciation research. Indeed, Catford coined the term "threshold of intelligibility" as early as 1950 (as cited in Nelson, 1992), and Gimson (1980) speaks of "minimal general intelligibility" or the lowest requirement for efficiently conveying a message from a native speaking listener's standpoint. Yet it is not clear why Morley placed the two communicative threshold levels where she did in her index, nor what these designations entail. Clearly, if the rating scale is to be widely adopted, or if it is to be used at a particular institution for a specific purpose, an empirical validation using speech samples from the appropriate population(s) would be desirable.[9] Indeed, Koren (1995), who views the pronunciation described in the literature as "unsatisfactory," advocates the need for standardized pronunciation tests, since current measures suffer from a low reliability. With empirical validation, Morley's index, which is a testable model, could raise the reliability bar considerably. See Turner and Upshur (2002) for problems with theory-based rating scales, however.

Here is a concrete example of why an empirical validation of Morley's index would make it more convincing. As described earlier, Morley directly relates intelligibility with its "impact on communication," which is described in terms of accent and its effect on the listener's perception. Yet Derwing and Munro (1997) and Munro and Derwing (1995a) have shown that the link between accent and intelligibility may not be that simple. Indeed, in two rater studies which explore the relationship between accentedness, comprehensibility, and intelligibility, what is unintelligible is almost always judged as being heavily accented whereas the opposite is not necessarily the case (i.e., what is heavily accented may or may not be unintelligible). Munro and Derwing's (1995a) study shows empirically that "foreign accent scores did not predict intelligibility

[9] Notably, the *Academic English Evaluation*, which was developed and is in use at the University of Michigan to assess the spoken English of students for their role as students, bases its intelligibility rating scale on Morley's index. Test-takers are also rated on functional language use and fluency (S.L. Briggs, personal communication, Sept 6, 2005).

very well" (p. 91), since accent was often rated more harshly than intelligibility and comprehensibility (a finding which was confirmed in Derwing et al., 1997), and that "the presence of a strong foreign accent does not necessarily result in reduced intelligibility or comprehensibility" (1995a, p. 91). In the same vein, Derwing and Munro (1997) conclude that "although some features of accent may be highly salient, they do not necessarily interfere with intelligibility. A clear implication of this finding is the need to disassociate accent ratings and intelligibility in language assessment instruments, which often confound the two dimensions" (pp. 11-12). This is a call for rating scales like Morley's *Speech intelligibility/ communicability index* to establish a stronger link between intelligibility and accentedness through empirical validation, in order to argue convincingly that scale descriptors in each column are grouped together appropriately.

In order to make sense of the above arguments, it is essential to examine how Derwing and Munro (1997) and Munro and Derwing (1995a) actually define their terms. As it happens, the authors, who are well aware of definitional ambiguities in the field, have done much to tease apart the constructs of "accentedness," "comprehensibility," and "intelligibility" by offering clear-cut, easily distinguishable definitions that are adhered to in all of their studies, albeit operationalized in different ways. Derwing et al. (1998) define "intelligibility" as the amount of utterance that the listener successfully processes,[10] in contrast to "comprehensibility," which is a more subjective judgment of ease of understanding the speech based on listener perception, and "accentedness," which is the extent to which non-native speech differs from the native speaker norm.

Notably, Munro and Derwing's definition of intelligibility, which is the most objective and easily quantified of the three terms cited above, is a far cry from the listener effort or irritation aspect latent in "comfortably intelligible" (Abercrombie,1956; Kenworthy, 1987) and in "functional intelligibility" (Morley, 1994). As it happens, Derwing and Munro's (1995a) "intelligibility" corresponds more closely to Smith's (1992) "intelligibility" than to other definitions of intelligibility that we have encountered thus far in the Literature Review. What is novel about Smith's definition of

[10] In Munro and Derwing (1995a), "intelligibility" is worded as the amount of "message" (i.e., not "utterance") that is "understood" (not "processed") by the listener. An analysis of the authors' work on the matter, however, leads one to conclude that this small variation in wording is negligible - although the way that the authors have conveyed the term may have evolved slightly, it still amounts to the same thing.

"intelligibility," however, is that it constitutes, along with two other terms that are featured in his chapter, different degrees of understanding on a continuum. He defines "intelligibility" as word or utterance recognition, "comprehensibility" as word or utterance meaning, and "interpretability" as the underlying meaning behind a word or utterance. Were these three terms to actually be mapped on a scale, then, intelligibility would be at the lowest degree of understanding and interpretability at the highest. However, as Atechi (2004) points out, it is not clear, using this categorization, at what point one category ends and the next one begins. In contrast, the relationship between Derwing et al.'s (1998) "intelligibility," "comprehensibility" and "accentedness," although unmistakably related, is not thought to be quite so linear.

To add to the confusion with regards to nomenclature, Derwing and Munro's (1995a) "intelligibility" is essentially equivalent to Gass and Varonis's "comprehensibility" (1984), which, as we will see, is operationalized in the same way as "intelligibility" in Derwing and Munro's (1997) study. Further, their term is similar to the objective interpretation of Smith's (1992) "comprehensibility," which is defined as "the degree to which the interlocutor understands what is said or written."[11] Smith, in fact, actually distinguishes "comprehensibility" from "irritability," noting that "while comprehensibility can be rated fairly objectively, irritability cannot" (p. 275). So too is it apparent that Derwing and Munro's "intelligibility" is much more readily quantifiable than the interpretations of "comfortably intelligible" that we have encountered and than the complex notion of intelligibility that plays into Morley's *Speech intelligibility/ communicability index* (1994).

In Morley's index, the intelligibility descriptor in band 3 exemplifies this complexity. The descriptor reads, "speech is reasonably intelligible, but significant listener effort is required because of the speaker's pronunciation or grammatical errors, which impede communication and distract the listener; there is an ongoing need for repetition and verification" (1994, p. 76). In the other band descriptors in the index however, errors in pronunciation or grammar are not mentioned at all, which points to an

[11] The subjective interpretation of Smith's term would presumably coincide more closely with Derwing and Munro's "comprehensibility."

inconsistency in the rating scheme and makes any attempt to measure or even define "intelligibility" difficult.

In light of this, Munro and Derwing (1995a) have done much to get the definition of intelligibility down to its bare bones so that it can be readily quantified. Their definition is also much narrower than Fayer and Krasinski's "intelligibility" (1987), which, as "one aspect of the total communicative effect of a nonnative message," has both linguistic and non-linguistic sources (p. 313).[12] Indeed, Dalton and Seidlhofer (1994), who state that the goal that should be adopted for learners is "comfortable intelligibility" (borrowed from Kenworthy, 1987) or acquiring an intelligible accent, emphasize that intelligibility, far from being limited to linguistics, is often overridden by economic and cultural factors and is often linked to issues of language identity.[13] Moreover, Fayer and Krasinski (1987) remark that negative attitudes towards speakers of a particular variety of English tend to also decrease intelligibility in the ears of the listener. This notwithstanding, the present study will only focus on intelligibility as it relates to second language pronunciation, which is a difficult task in itself. (For a sociolinguistic perspective on accented English, see Gatbonton et al., 2005; Eisenhower, 2002).

Measuring Intelligibility

We have seen from the above that there is no universal consensus on a definition of intelligibility. Perhaps it is logical, then, that there is also no "universally accepted way" of assessing intelligibility (Munro & Derwing, 1995a, p. 76) and that none of the methods that have been used can be said to be completely satisfactory (Brown, 1989). This makes it all the more interesting, perhaps, to see how intelligibility, as it has been differentially defined in the literature, has been played out in different empirical studies.

One clear-cut way to measure intelligibility in terms of the amount of message understood by the listener (Munro & Derwing, 1995a) is to actually get the listener to write down exactly what is heard and then quantify the amount of each non-native speaker's message that the listener was able to decipher. In Gass and Varonis (1984),

[12] Notably, in stating that "intelligibility is hearer-based; it is a judgment made by the listener," Fayer and Krasinski place most of the onus in deciphering the message on the listener (1987, p. 313).

[13] Looking through a Derwing and Munro-inspired lens, it becomes evident that Dalton and Seidlhofer (1994) do not make the same distinction between "accentedness" and "intelligibility."

"comprehensibility" (which we will remember can be roughly equated to Munro and Derwing's "intelligibility") was measured in this way, as native English speaking listeners orthographically transcribed sentences read aloud by non-native speakers of English. Scores were then assigned based on discrepencies between the sentence that the speaker produced as recorded in the transcriptions and the actual story scripts from which the sentences were drawn.[14]

Derwing et al. (1997), Derwing and Munro (1997), and Munro and Derwing (1995a) have also operationalized intelligibility using native English speakers' orthographic sentence transcriptions of non-native speech (i.e., dictations), whereby the transcriptions are coded using exact word-matching. Derwing et al. acknowledge that while this is perhaps a conservative way of measuring intelligibility as compared with other methods, it does have the advantage of being simple and objective (1997). In the three studies cited above, this measure of intelligibility is juxtaposed with listener judgments of comprehensibility and/ or accentedness using a 9-point rating scale. It is clear that the operationalization of these terms is consistent with the authors' claim that intelligibility is the least subjective of the three measures.

Another way of measuring intelligibility is through use of a cloze test. In a study on world Englishes, Smith (1992) measured the intelligibility of both native and non-native varieties of English by giving listeners (both native and non-native) a fixed cloze test. Intelligibility scores were calculated by tabulating the number of blanks that the listener was able to fill in.

Intelligibility can also be measured subjectively or impressionistically. Indeed, Kenworthy (1987) suggests that the easiest way to assess the intelligibility of particular speakers is to simply ask a listener how easy or difficult they are to understand. This method, it should be noted, corresponds much more closely to Derwing et al.'s "comprehensibility" than to their notion of "intelligibility" on account of the listener's greater interpretative scope (1998). Kenworthy also argues that having listeners rank order non-native speakers for intelligibility has been shown to be consistent with more objective assessments of intelligibility, although she doesn't cite any studies to back this claim.

[14] Missing or incorrect words, for example, were counted as errors.

Fayer and Krasinski (1987) show one way of how impressionistic ratings of intelligibility might be implemented in an empirical study. After a first listening of a speech sample by a native or non-native speaker of English, native speaking listeners were asked to make an overall intelligibility judgment using a 5-point scale. After the second listening, they were asked to rate each speaker for some of the features that seem to play into the authors' broad definition of intelligibility (e.g., grammar, voice quality, lexical errors) on separate 5-point scales. The authors contend that "factors affecting intelligibility are complex" (p. 314).

Another way of measuring intelligibility impressionistically is to get listeners to mark on a rating scale grid how intelligible they find a given speech sample to be. Anderson-Hsieh et al. (1992) present such a rating scale to their trained raters in what the authors refer to as "ratings of pronunciation." The lowest point of the 7-point scale represented "heavily accented speech that was unintelligible," the midpoint represented "accented but intelligible speech," and the highest point "near native-like speech" (1992, p. 538).[15] Of course, when intelligibility is measured impressionistically on a rating scale, the assumption is that it is a scalar phenomenon (i.e., from more intelligible to less intelligible or, in the case of comfortable intelligibility, from easily intelligible to intelligible only with an insurmountable difficulty) rather than a binary, all-or-nothing phenomenon (i.e., intelligible vs. unintelligible). (See Brown, 1989).

In sum, although there are many different ways to measure intelligibility, the basic choice is whether to measure objectively or impressionistically (i.e., subjectively). It should be reiterated there is no best way to measure intelligibility – each method has its drawbacks. Effort should be made on the part of the researcher, however, to ensure that the theoretical definition of the construct and the way it is operationally defined are as congruent as possible.

[15] As the rating descriptors make plain, a methodological weakness of this study stems from its failure to separate accent from intelligibility. In addition, it should be noted that these impressionistic ratings conform more to Derwing et al.'s (1998) "comprehensibility" than to "intelligibility" because of the dependence on the listener's perceptions. In fact, in the literature review of their 1997 article, Derwing and Munro state, in reference to Anderson-Hsieh et al. (1992), that "unfortunately, these authors did not actually measure intelligibility" (p. 3). From this, it seems that Derwing and Munro's clear-cut yet exclusionary definition of intelligibility virtually precludes operationalizing the term impressionistically, since this by definition entails a larger degree of subjectivity that treads on comprehensibility's terrain.

Honing in on Pronunciation Features Critical for Intelligibility

The challenges and difficulties in defining and measuring intelligibility do not end here. Indeed, as Munro and Derwing state, "not only is there little empirical evidence regarding the role of pronunciation in determining intelligibility, but also there is no clear indication as to which specific aspects of pronunciation are most crucial for intelligibility" (1995a, p. 76). If intelligibility is to be (or has become) the new goal of pronunciation teaching, then the above is extremely problematic. Knowledge of which pronunciation features are the most critical for a speaker's intelligibility could prove helpful in making informed pedagogical decisions about what aspects of pronunciation to focused on in pronunciation instruction. If researchers cannot provide empirical evidence in what is "arguably the most pressing issue in L2 pronunciation research" (Field, 2005, p. 399), then pronunciation pedagogues will continue waving about in the air blindfolded. Anecdotal evidence or theoretical assumptions about what is important to focus on is simply not enough if intelligibility, as the goal of pronunciation teaching, is to attain some sort of legitimacy.

Nowhere does this become more apparent than in looking at the age old segmental-suprasegmental debate that is still pervasive in the field. (See Jenkins, 2000). In traditional pronunciation teaching, the focus of instruction was largely on "segmentals," or individual sounds units (i.e., articulatory phonetics) rather than "suprasegmentals," or features of pronunciation that span beyond individual sound segments (e.g., stress, rhythm, intonation). Consequently, thoughts on the relative importance of segmentals and suprasegmentals in pronunciation instruction were also extended to views on intelligibility, especially as intelligibility became more and more widely accepted as *the* goal of pronunciation. On the segmental side of things, Prator (1967) points to "phonetic abnormalities" and departures from phonetic or phonemic norms of the language as being very often (but not always) the cause of unintelligibility (p. xv). Notably, suprasegmentals are not mentioned in his discussion.

Many English language pedagogues have come to regard the teaching of segmentals as obsolete, prosaic, and ineffectual, symbolizing the worst of decontextualized pronunciation teaching (see Yule, 1990),[16] although Celce-Murcia et al. foresee a more balanced view towards segmentals and suprasegmentals in the making (1996). The claims, however, that "suprasegmentals are just as important as segmentals, if not more so, for achieving the objective of intelligibility" (Rogerson & Gilbert, 1990, p. viii) or that that "stress, rhythm, and melody of English words and discourse" are "those features of English pronunciation that affect intelligibility the most" (Hahn & Dickerson (1999, p. 1), are, it seems, unsubstantiated by empirical evidence.[17] Indeed, Anderson-Hsieh stresses that the results of several empirical studies "are only suggestive rather than strongly conclusive of the greater influence of suprasegmentals on intelligibility" (1995, p. 17), and Hahn contends that, while assertions about the value of teaching suprasegmentals and relationship between suprasegmentals and intelligibility are based on "a theoretical understanding of prosody in discourse, they offer little if any empirical evidence to support (their) claims about how suprasegmentals affect intelligibility" (2004, p. 203).

In an recent study that set out to identify those pronunciation features which are critical to intelligibility, Raux and Kawahara (2002) investigated the relationship between ten pre-selected segmental and prosodic pronunciation errors common to Japanese learners of English and a linguist's intelligibility ratings. To measure the independent variable, participants read a diagnostic passage, and Automatic Speech Recognition software computed error rates for each error type. Results show that errors such as non-reduced vowels and vowel insertion, which are related to sentence rhythm and word stress, were found to be more crucial to intelligibility than the strictly segmental errors. This suggests that the role of suprasegmentals in intelligibility may be superordinate to that of segmentals, but replications are clearly necessary to establish this pattern. Furthermore, the study operates on the assumption that as error rates decrease intelligibility rates increase, whereas the link between error frequency and intelligibility has yet to be established empirically (although it seems to make sense intuitively).

[16] It should be noted that a focus on segmentals in pronunciation pedagogy is still quite prevalent in EFL settings.

[17] Note that these two quotations are taken from actual pedagogical materials which presumably must sell their method.

Intelligibility and the ITA Context

Several other empirical studies which attempt to determine those features of pronunciation which are most essential for intelligibility are part of a growing body of literature that addresses what has been referred to as the "international teaching assistant (ITA) problem" (Hoekje, & Williams, 1994) or "foreign TA problem" (Mendenhall, 1996), namely that ITAs coming from a different educational system often have difficulty adapting to the North American instructional context (Bauer, 1996); that ITAs often have trouble communicating well with their undergraduate students (Tyler, 1992); that undergraduate students, who have been charted for their lack of receptivity to ITAs (Mendehall, 1996), often express negative attitudes about ITAs' effectiveness as instructors (Johncock, 1991; Oppenheim, 1998). ITA programs must work to address some of these problems, particularly because of attempts of some states "to take up legislation to hold universities accountable for ensuring that college instructors be verbally understandable to their students" on the one hand (Mendenhall, 1996, p. 232), and because of ITAs' rights to file legal claims for civil rights violations (i.e., as victims of discrimination) on the other. (For more on ITAs and the American legal system, see Oppenheim, 1997). Thus, the training and assessment of ITAs, particularly in the United States, has become an issue of utmost political sensitivity (Johncock, 1991), and university administrators should be mindful of the legal ramifications (Oppenheim, 1998).

Preparatory courses for ITAs are a way for institutions to ensure that the ITAs do meet certain oral English language competency standards that have been set by some states (Oppenheim, 1998). Several of these courses, some of which are for graduate student non-native English speakers at large, and others of which are specially catered to the communicative needs of ITAs, have considered intelligibility to be one of the main speaking objectives. (See Graham & Picklo, 1994 for an example of the former and Smith, 1994 for an example of the latter). Thus, explorations of features of pronunciation that are critical for intelligibility have had particular resonance in ITA studies. Conversely, some ITA-related studies have explored which features of pronunciation are necessary for intelligibility. Hahn (2004) and Anderson-Hsieh et al. (1992), referred to earlier in this chapter, are two examples of such studies.

In an attempt to strengthen claims about the pedagogical value of teaching suprasegmentals in the pronunciation literature (see earlier in the chapter), Hahn (2004) provides empirical evidence that primary stress significantly affects the discourse of non-native speakers of English. A native Korean speaking ITA with a high spoken English proficiency and graduate level training in phonetics was recorded reading the text script of an ITA lecture. Ninety monolingual undergraduate students were then randomly assigned to three experimental conditions, the groups differing only in terms of which artificially manipulated version of the ITA's recording they listened to. In the first group (native-like condition), normal primary stress placement was retained to contrast new and given information; in the second and third groups (non-native like conditions), primary stress was misplaced and absent respectively. The listeners had to demonstrate how much they had understood by first writing down as much as they could remember from the lecture and, second, answering short answer comprehension questions. The first group outperformed the other two groups on all measures, (although this was not statistically significant for the short answer questions, possibly due to the instrument not being refined enough). The first group also wrote more consistently positive comments about the speaker than the participants in the other groups, whose comments tended to be more mixed.

This experimental study sets an important precedent for future studies (e.g., Field, 2005) in that it systematically isolates one feature of discourse and pronunciation in order to gauge its relationship with intelligibility in a controlled setting. It also offers insight into how native speakers, and in particular undergraduate students, react to systematic variations in non-native speech.

In an earlier study which compares non-native English speech "deviance" (i.e., errors) in the areas of segmentals, prosody, and syllable structure with impressionistic pronunciation judgments of native-speaking raters, Anderson-Hsieh et al. (1992) provide evidence that prosodic variables, in fact, have a stronger effect on intelligible pronunciation than do either speech segments or syllable structures, although they were all found to have a significant influence on pronunciation ratings. There are a great deal more variables at work here than in the aforementioned study (Hahn, 2004), and the approach is wholly different, but the study is executed with care and the result is, likewise, a step towards the empirical validation of something we know little about.

The speech samples utilized in the Anderson-Hsieh et al. (1992) study are recordings of reading passages from the *SPEAK* test, a retired form of the *Test of Spoken English* (*TSE*) that is widely used at institutions of higher learning in North America to evaluate the spoken English of ITAs, often for screening purposes (Celce-Murcia et al., 1996).[18] The fact that the three raters in the study were ESL teachers who had actual experience rating the *SPEAK* test (and thus by implication had already made decisions about spoken proficiency based on recorded *SPEAK* test samples in a real university context) lends authenticity to the study and makes the findings all the more resonant for the ITA context. Further, the authors explicitly link the construct of intelligibility to the pronunciation subtest of the *SPEAK* test, claiming that "the criteria used for pronunciation in the *SPEAK* test are based on considerations of intelligibility and acceptability,[19] and the raters are instructed to judge errors mainly as they affect intelligibility"[20] (p. 530). Thus, the natural elaboration of this point is that what the raters were asked to do in this study is not so different than what they would actually do if they were rating using the pronunciation subtest of the *SPEAK* test, given the way they were required to listen to and evaluate the speech.

After listening to the speech samples a first time, the raters rated pronunciation impressionistically on the rating scale outlined earlier in the chapter (described in terms of intelligibility); after the second listening, they impressionistically rated a series of prosodic criteria, including stress, rhythm, intonation, phrasing, and overall prosody (an overall impression of the other 4 criteria) on a 4-point rating scale ranging from "least native-like" to "most native-like" (p. 541). These ratings were then correlated with each area of "deviance" that had been detected.

[18] In a survey of ITA training programs at American universities, the *TSE*/ *SPEAK* test was found to be the most common form of spoken assessment (Grove, 1994).

[19] See Ludwig (1982) for a definition of "acceptability."

[20] The authors describe that there are four subtests on the *SPEAK* test (and, therefore, on the *TSE*), including comprehensibility, pronunciation, grammar, and fluency. It should be noted that since the article was published, however, the *TSE* has undergone considerable revisions and the rating scale has been revised to reflect communicative competence. (For a history of the *TSE*, see Educational Testing Service, 2001; Caldwell & Samuel, 2001).

In order to ascertain what was deviant and what was not, speech samples of 3 native English speakers of American English reading the same passage were phonetically transcribed to establish a native-like norm. Then, recorded speech samples of 60 speakers representing 11 different language groups were also transcribed. The major error categories that were generated from analyzing the speech samples (again, with the native samples as a point of reference of what is deviant and what is not) were segments, syllable structure, and prosody.

While this study also provides empirical evidence that suprasegmentals play a greater role in intelligibility than segmentals, these results must be taken with a grain of salt due to methodological limitations (Anderson-Hsieh, 1995). Nor does the study suggest that the role of segmentals should be discounted, since segmentals too were shown to affect the pronunciation judgments. The time is ripe for more empirical studies which seek to identify those pronunciation factors that are key for intelligibility and those which are merely subordinate or superfluous to intelligibility (but maybe the cherry on the cake in terms of making the speech pleasant to listen to). Hopefully researchers will embark on this challenge and give to the field of pronunciation some badly needed empirical rigor.

To TSE *or Not to* TSE: *Is that the Question?*

As mentioned earlier in the chapter, the *TSE* and its institutional version, the *SPEAK* test, are the most widespread ITA assessment instruments in use across North American campuses. As a measure of global speaking performance, the rating scale of the 1995 version of the *TSE* cites "pronunciation" as one of many features to consider in rating, but leaves this construct largely unanalyzed (Educational Testing Service, 1995).[21] Clearly, there is room for a standardized assessment instrument which takes on a slightly more sophisticated view of pronunciation. (See Koren, 1995).

Yet many, like Tyler (1994), are of the opinion that "as the sole assessment for determining the readiness of a nonnative speaker to provide comprehensible academic discourse" the *SPEAK* test is simply inadequate (p. 727). Tyler bases this assertion on an

[21] In the 1995 version of the *TSE*, "pronunciation," which is defined as "the production of speech sounds" is listed along with "grammar," "fluency," and "vocabulary" as pertaining to "accuracy" (Educational Testing Service, 1995, p. A-8). Linguistic accuracy, in turn, is "only one of several aspects of language competence related to the effectiveness of oral communication" (p. 6).

empirical study which uses a qualitative discourse-analytic framework to analyze the spoken English of a native Chinese speaking graduate student. Despite the speaker's test score on the *SPEAK* test, which was reportedly at the almost-acceptable-to-TA cutoff mark set by the institution, the listeners, who were native English speaking graduate students in linguistics, had difficulty understanding the discourse. (One can only guess, then, how much undergraduates with no training in linguistics might have been able to understand). This suggests that the *SPEAK* test as the sole form of ITA assessment might not be enough.

While some North American institutions mandate that all non-native English speaking international students submit *TSE* or *SPEAK* test scores in addition to *TOEFL* scores for graduate school admission (though in time, the introduction of the speaking component of the new *TOEFL* is likely to render the *TSE* and *SPEAK* test obsolete), other institutions use the *TSE* or *SPEAK* test to assess the spoken English of prospective ITAs only (i.e., not all non-native English speaking graduate students at large). Still other institutions have developed their own ITA assessment instruments to be used in conjunction with the *TSE* or *SPEAK* test. (For a survey of assessment practices at American universities, see Johncock, 1991; Bauer & Tanner, 1994). These instruments, which can be used for diagnostic, screening, or placement purposes, are often used exclusively at the institution in which they were developed, though, as Smith (1994) states, it might be useful to eventually implement standardized ITA assessment instruments and scoring procedures across campuses. Until this (which does not look to be any time soon) occurs, however, most language proficiency entrance requirements and ITA language assessment information are readily accessible on university websites. Some American universities even detail their "International Teaching Assistant Policy" (University of Miami) or "Teaching Assistant English Language Proficiency Certification" (Temple University).[22]

As part of a policy to ensure that all instructors are proficient in English before they are assigned teaching responsibilities, the University of Illinois at Urbana-Champaign requires all non-native English speaking TAs to submit a *TSE* or *SPEAK* test

[22] This information was retrieved at the following URLs:
http://www.miami.edu/UMH/CDA/UMH_Main/1,1770,29045-1;39236-2;39234-2,00.html
http://www.temple.edu/ita/TA%20English%20Language%20Certification%206-1-2005.pdf

score in order to screen for oral proficiency.[23] In addition, the university senate mandates that all students (graduates and undergraduates) who have received below 610 on the paper-and-pencil *TOEFL* or 253 on the computer-based *TOEFL* (including prospective ITAs with these scores) take the Illinois *ESL Placement Test* (*EPT*), and some departments require even higher standards.[24] (Cutoff scores for the internet-based next generation *TOEFL* have not yet been announced).

The *EPT*, which was developed at the University of Illinois at Urbana-Champaign for use specifically at that institution, consists of both an oral interview and a written test. Of concern in this discussion is the oral interview, which consists of two parts. The first part, called the "Global Proficiency Assessment," adopts intelligibility as its central measure; the second part, called the "Content-Specific Assessment," adopts accuracy in speaking as its central measure.

The Global Proficiency Assessment, which takes approximately 5 minutes to complete, also has a binary set-up. The first part consists of the interviewer asking the interviewee questions about the form that was filled out. This functions both to ensure the accuracy of the information on the form, and to give participants an opportunity to listen to one other. In the second part of the Global Proficiency Assessment, the interviewee generates 3 minutes of unrehearsed, spontaneous speech based on a prompt on a topic that is not normally discussed in day-to-day life. The interviewer is instructed not to turn this into a conversation so as to maintain focus on the central questions about the speaker's intelligibility.

The first question that the interviewer must answer about the speaker's intelligibility is, "can I understand every word that the interviewee says?" The word "understand" is taken to mean that the interviewer comprehends each word immediately so as not to have to guess at words, and does not need to wait for additional context in order to derive meaning about what has been said. The second question that the interviewer must answer is, "Is there any evidence that the interviewee misunderstood

[23] Notably, this policy applies not only to prospective TAs who are "international" (i.e., ITAs), but also to U.S. born citizens, passport holders, and permanent residents for whom English is not a first language. This information was found on http://www.provost.uiuc.edu/provost/announce/oralenglishpolicy.htm

[24] This information was generously provided by Wayne Dickerson (personal communication, March 31, 2005), professor at the University of Illinois at Urbana-Champaign and developer of the oral interview of the *EPT*. Much of it can also be found on the website: http://www.deil.uiuc.edu/about/ept.html

anything that I said?" Non verbal cues, long pauses, and repetition requests on the part of the interviewee can constitute evidence of this.

When Wayne Dickerson, who developed the oral interview of the *EPT* test, was asked why intelligibility was selected as the criterion for the Global Proficiency Assessment, he replied:

> The whole purpose behind testing students for their English proficiency in the first place is to determine if they have a strong enough competence in English to carry academic work at this institution. For the Oral Interview, the criterion is two-way intelligibility: It is a measure of clarity of pronunciation and accuracy of perception, two components deemed to be key in the student's likely academic success (personal communication, March 31, 2005)

This conception of "two-way intelligibility" exposes an idea which is fundamental to any mention of intelligibility, namely that there is a message sender and a message receiver. This brings to mind Morley's memorable quote that "intelligibility may be as much in the mind of the listener as in the mouth of the speaker" (1991, p. 499). Of course, in the case of the *EPT*, only the interviewer is making a judgment about intelligibility, but the notion of "two-way intelligibility" is still there. Dickerson's quote also unearths the idea that two-way intelligibility is essential for a student's (graduate or undergraduate) academic success given the nature of the tasks that must be carried out in the academic domain.

In the *EPT* oral interview, if the interviewer is able to respond to both intelligibility questions in the affirmative with no hesitation, then the oral interview is discontinued and the interviewee is exempt from pronunciation work in the ESL speaking course. If the interviewer does detect some sort of misunderstanding from one direction or the other, then one of two things can happen. If the misunderstanding was judged to be so great and the speaking so difficult, the interviewer may require the student to take the ESL oral course without any further testing. If, however, as in the vast majority of cases of students who take the test, the misunderstandings are intermittent, then the interviewer moves into the second part of the oral interview, the Content Specific Assessment, to see how well the interviewee controls the content of the ESL oral course that he/ she will be required to take. It is important to note that the *EPT* is a placement test and does not serve a diagnostic function.

The rich tradition of oral language testing at the University of Illinois at Urbana-Champaign, which can be traced back to the mid 1950s,[25] provides a striking contrast to the language assessment practices at McGill University, which has never, to my knowledge, developed a university-wide English language oral proficiency test, and does not require its non-native English speaking graduate students to take any form of spoken assessment for either admission or placement purposes.[26] In fact, the *TSE* (*Test of Spoken English*) and *SPEAK* test do not come up at all in a search through McGill web pages.

Non-Canadian graduate student applicants to McGill University who do not speak English as a first language and do not have an undergraduate degree from an institution where English is the language of instruction are required to submit either *Test of English as a Foreign Language* (*TOEFL*) or *International English Language Testing System* (*IELTS*) scores. Notably, this requirement does not apply to Canadian applicants whose first language is either French (i.e., Francophones) or a language other than English or French (i.e., Allophones). Indeed, by virtue of being Canadian citizens, these applicants are absolved of all proof of English language proficiency requirements even though there is no guarantee that their oral language skills meet the language standards to which international students are held.

While the introduction of the next generation *TOEFL* in September, 2005 will introduce a speaking component to standardized English language testing, this does not reduce the need to have some additional form of developed oral assessment (which could conceivably include "intelligibility") for all non-native English speaking students to ensure that they have the oral language skills (and the concomitant intelligibility) they need to carry out the tasks that are required of them, and particularly for ITAs to ensure that their spoken language proficiency (and intelligibility) does not preclude their ability to carry out instructional tasks. Further, if it is true that "pronunciation is the most overtly identified problem associated with ITAs" (Hoekje & Williams, 1994, p. 14), then this "problem" needs to be addressed in both ITA assessment and ITA training programs, neither of which exist per se at McGill.

[25] Lyle Bachman, Wayne Dickerson, and Fred Davidson play a part in the history and evolution of English language testing at the University of Illinois, Urbana-Champaign.

[26] This being said, non-native English speaking graduate students do have the option of enrolling in the *Communication and Pronunciation for Graduate Students* course offered through the McGill English and French Language Center.

Final Thought

The lack of a universal definition of intelligibility, of a field-wide consensus for the "best way" to assess intelligibility, and of an empirical basis for identifying which features of pronunciation are the most critical for intelligibility have rendered the construct of "intelligibility" problematic despite the widespread use of the term. These gaps in our knowledge, which have been elaborated in this review of the literature, hinder our ability to both determine which aspects of pronunciation should be focused on in the second language classroom (if intelligibility is, indeed, to be the goal of second language instruction), and to reliably assess intelligibility for our intended purposes. The present study attempts to address some these issues in a context in which the stakes for intelligibility are thought to be high - the ITA context.

Chapter 3: Method

Statement of Purpose

This descriptive, mixed-design study, which explores the interface between pronunciation and second language assessment, treads on relatively untrodden ground. The primary purpose of the study is to examine the validity of "intelligibility" as a criterion for assessing oral proficiency in the pronunciation of non-native English speaking graduate students; the secondary purpose is to identify those features of pronunciation which are most crucial for intelligible speech.

Part of the drive which fuels this study and the pursuit of the research questions is the recognition for the need to pull pronunciation assessment out of its state of "neglect" (Goodwin et al., 1994) and invigorate it by drawing on underlying educational realities. The study is likely to raise issues which are of current relevance to university administrators, professors, graduate students, undergraduate students, and pronunciation and assessment specialists alike.

Research Questions

This study explores whether intelligibility is an appropriate criterion for assessing the oral proficiency of non-native English speaking graduate students in the academic domain. The primary research question addresses whether intelligibility is "enough," that is, a sufficient goal and an adequate assessment criterion, for evaluating the pronunciation of non-native English speaking graduate students in the academic domain. If intelligibility is deemed to be "enough," then is there a threshold level (i.e., a minimum acceptable level) of intelligibility that can be identified? If not, then what pronunciation criterion might be more suitable?

Irrespective of the answers to the above queries, the secondary research question seeks to shed some light on those features of pronunciation which are most crucial for intelligibility in an effort to determine what constitutes intelligible pronunciation.

In the Results chapter of this thesis, the two main research questions, stated succinctly below, will be answered as follows:

1. Is intelligibility "enough," that is, a sufficient goal and an adequate assessment criterion for evaluating proficiency in the pronunciation of non-native English speaking graduate students in the academic domain?

2.	Which features of pronunciation are most crucial for intelligibility?

Research Participants in the McGill University Context

Definition of Native versus Non-Native English Speaker

The participants consisted of nineteen non-native English speaking graduate students and 18 native English speaking undergraduate students at McGill University in Montreal, Canada. Participant profiles and recruitment procedures will be detailed later on in the chapter.

Particularly in a context as linguistically diverse as Montreal, Canada, where bilingual and multilingual individuals abound and the lines between first language and second language are often blurred, it is essential to clearly define the terms "native speaker" and "non-native speaker" in a way that is transparent to prospective participants. In fact, in a few grey area cases, the definition of these terms directly affected eligibility to participate in the study. For the purposes of the study, a "native speaker" was defined as someone who had had English at home before the age of 3, whereas a "non-native speaker" is someone had had no such exposure to English at that early age.

The rationale for the choice of age 3 is that it is commonly held among speech specialists that speech and language development is at its most intensive during the first 3 years of life. (See, for example, the NIDOCD website, 2005).[27] Furthermore, there is little consensus in SLA research as to when exactly in a person's maturational development (i.e., at what "critical" age) it may no longer be possible to acquire native-like pronunciation in a second language (Han, 2004). Research has, however, shown that children are far more adept at acquiring native-like command of a second language phonological system than adults (Scovel, 2000). Whereas at age 13, it is highly unclear whether or not phonological acquisition might become constrained by a loss of plasticity in the brain, thus resulting in a "foreign accent," at age 3, the acquisition of native-like pronunciation is far less contested. (See Celce-Murcia et al., 1996; McLaughlin, 1978). Thus, for the purposes of this study, age 3 was thought to be a "safe" cutoff point for the

[27] NIDOCD stands for the National Institute on Deafness and Other Communication Disorders. It is part of the National Institute of Health (NIH) which, in turn, is part of the U.S. Department of Health and Human Services.

distinction between native and non-native speaker, far removed from most researchers' demarcation of the critical period, and, therefore, free of any unnecessary ambiguities.

The Speakers

The 19 non-native English participants (6 male, 13 female) were all full-time graduate students in the Faculty of Education at McGill University, ranging in age from 24 to 42 years (*M*=28). They were from a variety of L1 backgrounds, including: Japanese (5), Mandarin (5, including a Mandarin-Taiwanese early bilingual), Korean (3), and one of Malay, Sundanese, Bahasa Indonesian and Javanese (early bilingual), Argentinean Spanish, Serbo-Croatian, and Quebec French respectively.[28] As a collective, the group will be referred to as the "speakers," and in referring to an individual speaker, the pseudonyms "Speaker A" through "Speaker S" will be used. Descriptive speaker data can be referenced in Appendix H.

All speakers reported taking either the *TOEFL* or the *IELTS* with the exception of Speaker R, the native speaker of Quebec French, who as a Canadian is exempt from English language proficiency tests for university admission. In the 2004-2005 academic year, the minimum *TOEFL* score for admission into Integrated Studies in Education, the department to which 16 of the 19 speakers were accepted, was 580 on the paper and pencil test or 237 on the computer-based test.[29] The 3 Indonesian participants, who submitted *IELTS* scores, needed an overall minimum band score of 6.5 to get into the program – the same as the McGill minimum.[30] That all speakers in this sample met and often surpassed the minimum scores on these standardized tests speaks to the fact that they have obtained a level of competency in English that is thought to be acceptable by both the university and the department, since no additional language testing is required of them.

Of the 18 speakers who reported having teaching experience, 13 had taught English as a Second or Foreign Language, 3 had had experience as Teaching Assistants, and 2 of the 3 were actually employed as TAs during the time of the data collection. As

[28] The online edition of *Ethnologue* was consulted to find out information on language families and, in particular, spoken languages in Indonesia (Gordon, 2005).

[29] The McGill minimum score for that same year was 550 on the paper and pencil test and 213 on the computer-based test. Individual departments can choose to set higher cutoff scores than the university.

[30] According to Carolyn Turner, the current Director of Graduate Programs in the Faculty of Education, submission of *IELTS* scores is a rare occurrence in the department.

for what they envisioned after graduating, 11 indicated that their career plans included English language teaching and 5 were intent on pursuing academic careers.[31]

It should be noted that only a select number of graduate students who were in contact with me through a graduate course (convenience sample) or who had heard about the study from a fellow student through word of mouth (snowball sample) were invited to participate in the study, and of those students only the ones that volunteered and could be scheduled actually participated. The process of finding participants in this study alerted me to the practical difficulties of obtaining a sample that is truly random in educational research, although a random sample was not even attempted here. Indeed, as my decisions later on in the study will show, obtaining a sample with as much L1 diversity as possible was a much larger concern than either random sampling or sample size.

The Raters

Native English speaking undergraduate science students enrolled in the summer course, *Topics in Organic Chemistry*, were invited to participate in this study as raters of the speech samples. A total of 18 students, ranging in age from 18 to 24 years (*M*=20), participated in four rating sessions. Half were in their 2nd year of studies, 4 were in their and 1st and 3rd years respectively, and 1 participant was a 4th year student. None of the raters reported ever taking a linguistics course or having training in phonetics, phonology, or anything related to pronunciation. In that sense, they truly are "untrained" or "naïve" raters.

All raters function predominantly in English, with 15 indicating that they speak English 100% of the time and 3 that they speak English 75% of the time. This is perhaps unsurprising given that these raters, although in a French speaking province, attend an English speaking university with a strongly Anglophone culture in a multilingual city. Nonetheless, half of them report having some knowledge of a language other than English, and 1/6 of the sample purportedly have knowledge of two languages other than English, although it is unclear what their proficiency level is. The remaining 1/3 of the sample are monolingual English speakers.

[31] These categories are not mutually exclusive.

Instruments

Speech Data Elicitation Instrument

As discussed in the previous chapter, there exists, as yet, no standardized pronunciation test, and those instruments that do exist fail to meet high standards of reliability and validity (Koren, 1995). Nor is there any universal consensus on how to measure or even define intelligibility. These considerations presented problems in determining what kind of instrument to use to elicit speech data for this study: whether to construct a speech elicitation instrument for the purposes of this study when there is no viable model to follow, or whether, instead, to opt for a "canned instrument," when the implication of using such an instrument is that those tasks are sufficient for the purpose of assessing proficiency in pronunciation.

Evidently, there are trade-offs with both options. What is clear from language assessment and pronunciation literature, though, is that task characteristics do affect a test-taker's performance (Bachman & Palmer, 1996; Celce-Murica et al., 1996; Briggs, 1994). For example, a speaker's performance in reading a diagnostic passage (which is crafted to elicit vowel sounds and consonant clusters that would not necessarily arise in spontaneous speech) is likely to be different than performance on a task which elicits free speech. Notably, in the former task, the intervening variable of reading ability must be contended with. (See Munro & Derwing, 1994 for a study which attempts to eliminate some of the differences between reading and speaking conditions). Perhaps it can be understood, then, why Koren, in her cry for guidelines for a "greatly improved pronunciation test," highlights the importance of capturing different types of speech situations (p. 390). On a similar note, Celce-Murcia et al. (1996) suggest that spoken production samples be obtained for both types of tasks at the outset of classroom instruction for diagnostic purposes. This would yield a more rounded speech profile than complete reliance on one task.

Taking into account all of these issues, it was decided that a 1995 version of the *Test of Spoken English* (*TSE*)[32] would be used to elicit speech samples from the non-native English speaking graduate participants, since this instrument is commonly used at institutions of higher learning to assess the "ability of nonnative English speaking

[32] Reproduced by permission of Educational Testing Service, the copyright owner. See Appendix F.

graduate students to communicate orally in English" and, in particular, to screen ITAs (Educational Testing Service, 1995). Typical of a high-stakes standardized test, it has also undergone rigorous empirical validation (Educational Testing Service, 2001) and has achieved higher levels of reliability across test items than would be possible using a "home-developed" instrument, while, at the same time offering the considerable task variety called for by Koren (1995). It should be noted, however, that the *TSE* was used exclusively for the purposes of obtaining speech data for this study. The *TSE* rating scale was not used to evaluate the speech samples.

Non-Native Speaker Questionnaire and Interview Questions

A questionnaire for the non-native English speaking graduate students (i.e., the "speakers") was developed for the purpose of finding out information about their "background and goals as they relate to language and pronunciation." The questionnaire can be referenced in Appendix D,[33] and some of the data that were generated can be referenced in the table in Appendix H. This instrument was used in conjunction with the following three scripted interview questions:

- Can you tell me about your teaching experience?
- Have you ever worked as a Teaching Assistant (TA)?
- What do you plan to do after you graduate?

Details on the administration of the questionnaire and interview questions will be outlined later on in the chapter.

Native Speaker Questionnaire and Rating Scale

The Native Speaker Questionnaire and Rating Scale, which was specifically crafted for the undergraduate raters, generated qualitative and quantitative rater data for the purpose of shedding light on the two main research questions. The instrument, which is in three sections, can be referenced in Appendix E. Section 1 unearths descriptive data about the raters' age, language background and program of study. Section 2 elicits intelligibility ratings for each speaker, rankings on any pronunciation features which may have hindered the speaker's intelligibility, a decision as to whether the speaker's

[33] In the data analysis, the responses "to get your message across" and "to be understood" were collapsed to contrast with "to sound native-like" in question 13.

pronunciation is adequate for him/ her to TA an undergraduate course, and general comments about speech. Section 3 asks raters, at the end of the session, to identify those speakers that stand out as being easiest or most difficult to understand, and then to rank order the features of pronunciation which they surmise are most crucial for intelligibility.

In the instrument construction, which was informed by the analysis of the phonological data, it was considered of tantamount importance to provide both a clear-cut definition of intelligibility and a user-friendly rating scheme that would be readily accessible to the raters. The first part of the definition of intelligibility which is in use at the University of Illinois at Urbana Champaign from the *ESL Placement Test* (*EPT*) was adopted and adapted to this context (see Literature Review), and raters were instructed to mark approximately what percent of the speaker's words they were able to understand with an "X" on the scale from 0-100% that was provided. In a note above the rating scale, raters were reminded that "by understand I mean that you are able to comprehend each word immediately so you do not have to guess at words."[34]

The intelligibility ratings in this study, which are admittedly subjective and impressionistic, do not conform to the objectivity entailed in Derwing and Munro's more restrictive theoretical and operational definition of "intelligibility" (1997). However, the focus on intelligibility at the word level is reminiscent of Smith's (1992) definition or Kenworthy's (1987) operational definition. (See Literature Review). The first reason for this word-level focus was to encourage the raters to listen carefully to the individual words (and sounds) in assigning their ratings, in an attempt to divert their attention away from meaning or message. That is, the intent was to get them to shift focus from what was said (i.e., content) to the way in which it was articulated. Secondly, it was thought that the percent of *words* that are understood was more readily quantifiable, countable, and easier for science students to manage than the percent of *message* that is understood or anything having to do with meaning, which gets into fuzzier territory.

There is a comprehensibility component to this questionnaire which closely relates to Derwing and Munro's (1995a) "comprehensibility." Near the end of Section 2, the raters were asked, "How would you rate this speaker in terms of you being able to

[34] Note that the word "intelligibility" does not actually appear on the rating scheme, since it was felt that the construct could be perfectly explained without directly referring to it. This was a conscious decision in an effort not to complicate matters for the raters.

understand?" Raters then had four options to choose from, ranging from very easy to very difficult. Thus, in contrast to the intelligibility question, which relates to word level of speech, this question is about general ease of understanding and does not focus on word level.

It should be noted that in contrast to Derwing and Munro (1995a), who carefully distinguish between intelligibility and comprehensibility, this study simply assumes that the two constructs are closely related. In the Results section, the mean score for each speaker on the intelligibility scale is viewed as a check on the mean score of the comprehensibility question to try and gauge overall rating consistency.

Data Collection Procedures

Data Collection Phase 1: Speech Recording Sessions

All speech recording sessions for the non-native English speaking graduate students were conducted one-on-one in a quiet office in the Faculty of Education from March 22 to April 7, 2005. Speech sessions did not exceed 30 minutes, and the speakers received a remuneration of $10 for their time.

Having learned about the study through e-mail, by word of mouth, or in an announcement made in a graduate course, the speaker volunteers were required to fill out a Consent Form at the beginning of the recording session in accordance with McGill University ethical standards (see Appendix B), and any questions or concerns that they had about the procedures and purposes of the study were addressed. It was also emphasized to the speakers that the *Test of Spoken English* would be used to elicit speech data for the purposes of assessing their pronunciation only. In other words, they would not be judged on general English language proficiency nor on the content of their answers, as these considerations were not regarded as directly relevant for the purposes of the study.

Sound recordings, which had been piloted the week before, were made using *Sony Sound Forge Audio Studio 7.0* at 22,050 Hz, 16 bit, Stereo. After a quick warm-up, the speaker's microphone was adjusted. Speakers A and B used a handheld microphone, whereas Speakers C through S used headphones with an attached microphone.

In line with the actual administration of the *TSE* at a testing center, speakers were prompted by the *TSE* tape, which times responses to each item and thereby ensures standardization across participants. The *TSE* took less than 17 minutes to administer. Although I was present in the room, I had no interaction with the speaker during this time. At end of the session, participants filled out the Non-Native Speaker Questionnaire (Appendix D) and were asked the three scripted interview questions listed earlier in this chapter.

Once the data collection of the speech samples was complete, a lengthy process of transcribing and analyzing speech samples ensued. The process of *TSE* item selection for analysis and the procedures that were followed will be detailed later on in the chapter. For now, let us turn our attention to preparing the speech samples for the second phase of data collection, the rating section.

Interlude between Data Collection Phases

Stimulus preparation.

All four *TSE* items which had been transcribed into standard orthography (Items 2, 5, 7, and 12) for Speakers C-S were edited using *Cool Edit 2000, Sony Sound Forge 8.0,* and *Wave Pad 1.2.* Because Speakers A and B had been recorded using the handheld microphone rather than the microphone attached to a headphone, a set of extraneous sound variables were introduced which detracted from sample consistency. Thus, Speakers A and B were excluded from the editing process and were not considered as potential candidates in the rating session that was to follow.[35]

For speech samples of Speakers C-S inclusive, then, a series of steps were taken to improve the overall sound quality and account for objective and subjective measures of loudness. After reducing extraneous noise, the samples were normalized for absolute (objective) loudness by maximizing the volume to 98% without distortion. Then, the samples were adjusted for perceived (subjective) loudness by comparing a referent file (the sample which was thought to be the quietest) to a comparison file (all other samples). Peak and VU/PPM meters were consulted at this stage to ensure that there was no clipping in the files. Finally, the dynamic range compressor function was applied at the general voice level preset to make certain that the volume stayed within a prescribed range (threshold -20dB, ratio 4:1, limit 0dB).

[35] The unedited speech samples of all 19 speakers were, however, phonologically analyzed as will be discussed later on in the chapter.

These steps were taken to minimize differences in loudness across samples, or inter-sample loudness. Loudness *within* samples or intra-sample loudness, was not accounted for in the editing, however. That is, within each speech sample, the volume of the speech was not uniform, depending on such variables as speaker volume during articulation, the positioning of the microphone, the acoustic properties of the sound segments that were produced (e.g., fricatives are produced at higher frequencies than vowel sounds), etc. (See Rogers, 2000). As a result, there was considerable variability within samples, as is representative of normal speech (sometimes we speak louder, sometimes we speak quieter). All samples did, however, stay within the range prescribed by the preset in the dynamic range compressor.

Following a pilot session, the rating session was set at one hour. This included a teaching and practice component. Due to the potential of an order effect, different rating sessions were developed. Ordering by speaker was selected instead of ordering by item, since what was being sought was for raters to rate each of the speakers based on their general impression of intelligibility across the different tasks and not the speaker's differential performance on the different tasks.

Of the four *TSE* items that had been transcribed and edited (see later on in this chapter), Item 5, the story retelling task (60 seconds), and Item 7, expressing an opinion (60 seconds), were selected for the rating sessions because these tasks eliminated the intervening variable of reading ability that was a factor in other *TSE* items, and it was anticipated that these prompts would be quick for the raters to grasp. At the same time, the items were thought to be different enough from one another to present two distinct speech situations. (For arguments on the need to obtain different types of spoken production samples in pronunciation assessment, see Celce-Murcia, 1996; Koren, 1995).

Eight speakers, Speakers C, E, F, G, K, M, N and R, were "handpicked" for the rating session in an attempt to attain as varied as sample as possible in terms of L1 background and pronunciation proficiency (as informed by the analysis of the phonological data). The intent was to allow raters to be exposed to the widest possible gamut of speech in the rating section, since it was thought that this would be more instructive in answering the research questions than a sample that was purely random. Speakers G and N, it should be noted, were automatically assigned places among the 8 speakers since they were employed by McGill University as TAs in the 2004-2005 year. As it turns out, 4 male and 4 female speakers were chosen for the rating sessions. This was not deliberate and was only realized in retrospect.

A CD was prepared for each rating session, and the same "Practice Speaker," Speaker D, debuted at each of the sessions to give the raters a chance to go through the procedure once and become acquainted with the rating instrument. Then, the speech samples of the 8 speakers (the same speakers for all of the sessions) ensued in random order. In the *Questionnaire/ Rating Scheme*, speakers were labelled chronologically as "Practice Speaker," "Speaker 1," "Speaker 2," etc. based on the order of the speakers in that particular rating session. Items 5 and 7, played back to back, were subject to two listenings each so that, in the end, the raters listened to a maximum of 4 minutes of speech for each speaker in conjunction with filling out the instrument.

Data Collection Phase 2: Rating Sessions

Eighteen undergraduate raters participated in four different rating sessions held between July 14 and July 26, 2005. Five raters attended the first and last sessions and 4 raters attended the second and third sessions. The "treatment" that the different groups received was the same and the tracks were played at the same volume on the same CD player in the same room for each of the sessions. The only thing that differed across sessions was the order of the speakers, with the exclusion of the Practice Speaker, who was always Speaker D and always preceded the other speakers. Raters received a remuneration of $20 for 1 hour of their time.

Greeting the raters when they entered the room was a handout with the two *TSE* prompts on it and a packet containing the three sections of the questionnaire. Raters were alerted to the fact that, while Sections 1 and 3 would take little time to complete, Section 2 repeated itself no less than nine times (i.e., for the 8 speakers plus the Practice Speaker). After they had completed the first section, there was a 15 minute teaching session during which brief definitions of the pronunciation terms and illustrative examples (auditory, with visual reinforcement on the blackboard) for all the options that appeared in Section 2 were given. Then, for the "Practice Rating," raters listened to the Practice Speaker's picture task and the opinion task. At the end of this first listening (i.e., of both *TSE* items), they were asked to indicate on the scale the approximate percentage of the speaker's words that they were able to understand. Then, the same speech samples (two *TSE* items) were heard again and raters were asked to rank order a maximum of three pronunciation features that they felt hindered their understanding the speaker's words. If there were only two hindering features, then they were instructed to rank order only two; if none of the listed features prevented them from understanding words, they were instructed to simply leave it all blank and check the "none" box. It was emphasized that only those features that absolutely prevented the understanding of words should be identified. If something was noticeable or annoying but did not hinder understanding words, it should not be identified by rank ordering, but rather could be referred to in the comments. Similarly, any other feature that was not listed that hindered intelligibility or that the rater wanted to comment on, be it related or unrelated to pronunciation, could be simply indicated in the comments.

Of the pronunciation features they had rank ordered, raters were to check the most prominent problem only (i.e., just one of the two boxes for each identified feature). The same procedure was followed for all 8 speakers. To sum up, the first listening was a global listening, in which the raters were to focus on understanding the speaker's speech at the word level. If they did, perchance, find some words to be unintelligible, then during the second listening they were to play detective and try to identify what aspects of the speech may have caused the words to be unintelligible. This two-listening procedure, which moves from general to discrete pronunciation, is reminiscent of Anderson-Hsieh et al. (1992). (See Literature Review).

After the ratings of all 8 speakers were complete, raters progressed to Section 3, the "summing up" section, in which they identified any speakers that stood out in terms of being easiest or hardest to understand and listed (rank ordered) the top three pronunciation features which they felt had contributed most to intelligibility.

Data Analysis Procedures

Analysis of the Phonological Data

The purpose of the following section is to outline the data organization and preparation. The Overview presents a general summary of the procedure and is sufficient for interpreting the results in Chapter 4. Interested readers, however, may find the more detailed descriptions and, in particular, the process of developing the coding system insightful.

Overview.

The analysis of the phonological data, which will henceforth be referred to as "the phonological analysis," was an intensive, multi-step process that constituted a fine-grained analysis of the speakers' speech samples and culminated in a more holistic "intelligibility profile" for each speaker in the data set. The purpose was to investigate, in accordance with the research questions, whether and to what extent there was a loss of intelligibility in the speech samples, which pronunciation features might be responsible for any breakdowns in intelligibility, and, ultimately, whether intelligibility is "enough" for the graduate students in this sample. The unedited speech samples of all 19 speakers were phonologically analyzed, so as to be able to profile the pronunciation and intelligibility for all speakers in the sample. The sound quality of all recordings was deemed sufficient for this purpose.

The phonological analysis was initially bottom-up or data-driven. Efforts were made to transcribe the speech as accurately as possible such that problem areas would emerge just by looking at the transcriptions. The decision as to whether or not the speech was intelligible was suspended until the later stages of the analysis, when all of the segmental and suprasegmental transcriptions had been completed and the data had "spoken for itself," so to speak. Once the transcriptions were complete, color-coding was used in order to identify instances of unintelligibility and "unusual pronunciation."

In the second phase of the analysis, two theoretical models from Morley's (1994) article were applied to the data. This served the dual purpose of informing the analysis and providing a first step towards the empirical validation of these models.

Segmental and suprasegmental transcriptions.

The first step in the analytic process was to transcribe speech samples into standard orthography, a task that was facilitated through use of *Express Scribe 4.01*, an on-line transcriber. The following *TSE* items, which were thought to represent four distinct tasks, were selected for orthographic transcription: Item 2 – giving directions (30 seconds), Item 5 – story retelling (60 seconds), Item 7 – expressing an opinion (60 seconds), and Item 12 – presentation (90 seconds). (For the test prompts, see Appendix

F). Of these items, Item 7 is the most reminiscent of free speech and Item 12, with its written prompts, has elements of a diagnostic passage. The selection of these items conforms to Celce-Murcia et al.'s suggestion that free speech and a diagnostic passage should both figure into diagnostic testing (1996), and coincides with Koren's (1995) contention that a variety of speech situations need to be collected for assessment purposes.

TSE Item 12 was chosen for further phonological analysis, since the written prompts elicited many problematic sounds that did not come up in the other test items. In the Results chapter, the phonological analysis of Item 12 will be juxtaposed with the raters' quantitative and qualitative data of Items 5 and 7.[36]

Item 12 was phonetically transcribed using the IPA symbols referenced in both the *Handbook of the International Phonetic Association* (1999) and Pullum and Ladusaw (1996).[37] A narrow phonetic transcription was chosen over a broad phonemic transcription, since it was felt that this degree of detail would be necessary for the purpose of trying to detect intelligibility (or unintelligibility) in participants' speech. Diacritics, on the other hand, were overwhelmingly omitted, as they were not thought to be crucial in addressing the research questions. Thus, the word "plan," when pronounced as indicated in the Merriam-Webster on-line dictionary, was transcribed as [plæn] and not [pʰɫæn] (Merriam-Webster, 2005).[38] In cases where, for example, an unaspirated [p] (i.e., without the habitual puff of air at the beginning of the word) did, in fact hinder intelligibility, the [p] was simply color-coded accordingly when intelligibility judgments were made, to signal a problem with the articulation.

In the transcription of suprasegmentals, transcription conventions in Wennerstrom (2001) served as a guide, but ultimately, a notation system was developed, evolving as features "emerged from the data." Since suprasegmentals are often considered to be the "musical aspects of pronunciation" (Gilbert, 1994, p. 38), musical terms and concepts were included, where applicable. For instance, the Legend for Transcription Symbols

[36] This is not to say that the quantitative and qualitative analyses are in contrast to the phonological analysis, but rather that the ratings of the undergraduate student raters will be contrasted with the "pronunciation expert's" assessments of the speech samples in the next chapter.

[37] North American derivations from the IPA were not used in this study.

[38] Note that the dark, velarized [ɫ] was not contrasted with its light, lateral [l] allophone in the transcripts.

and Color-Coding in Appendix G makes use of the musical terms "tempo" (referring to the overall speed and pacing of speech), "staccato" (denoting sounds which are abruptly disjointed from one another), and "legato" (signifying smooth, connected speech where words flow seamlessly into one another). In addition, in instances of abrupt rises or falls that were perceived to be quite marked, when it was unclear whether these rapid fluctuations might have an impact on intelligibility,[39] some measure of the actual distance between pitches was desirable (i.e., how much the pitch went up or down). This situation at times arose for speakers in the sample whose native language is Mandarin, a tone language. The different tones used were difficult to show using the intonation markings (superscript and subscript) borrowed from Wennerstrom (2001), since high-medium-low register designations are too imprecise. Thus, in such instances, the musical interval between the pitches was calculated by determining the relative distance between the pitches.[40]

Below is an example one of the passages in which pitch fluctuations were notated. The excerpt is taken right from the end of Speaker I's response to Item 12, which ends in mid-sentence after she gets cut off by the "cassette examiner" (i.e., the next recorded *TSE* prompt), having used up her 90 seconds. Of the three lines, the bottom is the transcription in standard orthography, the middle line is the IPA transcription, with marks for primary stress, linking, and intonation where appropriate, and the top line shows which words the speaker emphasized (sentence stress) and, above the IPA transcription of "Professor Morgan," the pitch symbols. All coding and transcription conventions can be referenced in Appendix G.

We can see from the intonation markings that on the word "Professor," the speaker's pitch went up and then down and then down some more on the word "Morgan." The pitch markings more precisely indicate the rise in relative pitch by a Perfect 5th, down a Minor 3rd, and then down a Major 2nd.

[39] It will be remembered that decisions on intelligibility were suspended until all transcriptions were completed.

[40] Musical intervals are typically identified by their correspondence to a simple sound wavelength or frequency ratio (Lindley & Campbell, 2005), but musicians are often trained to approximate the distance between pitches by ear. My years of musical training greatly facilitated this process.

Figure 1

Pitch Change Notation in Transcriptions

```
                          ●                                 ● ↑P5    ↓-3 ↓+2
['ʧejnʤɜɪs to‿ðə‿'ʤʌʤɪs _ ʌ‿səw (1.) pɹə'fɛsəɹ 'mɔɹgən _ hæs
...changes to the judges uh so professor morgan has
          ●           ●
bin ri'plejst/ baj/ pɹə'fɛsəɹ]
been replaced by professor  (end of excerpt)
```

Color-coding for intelligibility.

Color-coding was used once the segmental and suprasegmental transcriptions were complete. A central challenge in this process was to find a standard against which to compare the speech samples in order to determine what is normative and what is not. The idea that the native-speaker standard, which stems from the traditional native-speaker/ non-native speaker dichotomy, is no longer an appropriate standard given the emergence of English as a global language, has permeated both the pronunciation literature (Jenkins, 2000), and the applied linguistics literature at large (Cook, 2002). In the phonological analysis of this study, therefore, these considerations were offset by an attempt to take into account the various "World Englishes" that the speakers indicated they had had exposure to in response to the questionnaire item that addressed this. (See question 10 in Appendix D). The phonetic symbols for different varieties of English (e.g., British, Australian) were referenced in Rogers (2000) when necessary.

Once the suprasegmental transcriptions were complete, data were color-coded for both "unusual pronunciation, but that does not affect intelligibility"[41] and "unintelligible pronunciation, or (pronunciation that) results in unintelligibility." (See Appendix G). Where possible, the incriminating pronunciation feature that was thought to be the cause of the "unusual pronunciation" or "unintelligibility" was itself color-coded. Unfortunately, the color-coding could not be reproduced in the printing of this thesis.[42] Thus, I will describe what was color-coded in Speaker A's excerpt below.

[41] The term "unusual" pronunciation was chosen in lieu of "deviant," "unnormative," or "accented," since it was found to be more neutral.

[42] Color-coded excerpts of the transcriptions can be obtained by e-mailing the author at talia.isaacs@elf.mcgill.ca

Figure 2

Coding for Intelligibility in Transcriptions

● ● ●

[sɔ‿plijs _ ʌ‿rimεm′bɜɹ/ »æn‿εks#$^{\underline{\text{biʃʌn}}\ \text{dejt}}$‿wɪl bi‿ɑlso‿

...so please uh remember and exbition date will be also

● ●! ●! ● ! ◇

ʧejndʒt« _ ʌ‿$^{\text{it}}$ _ $^{\text{it}}$ wʌz‿sʌpow′zd‿tu ɪt _ ʌ _ ði‿$^{\text{ˏɔɹɪʤɜɪnʌl ˎskεdzul}}$/]

changed uh it it was supposed to it uh the original schedule...

In the word [rimεm′bɜɹ], the displaced stress symbol ′ and unreduced vowel [ɜ] were color-coded in green for "unusual pronunciation," as were the unreduced [^] vowel, the indiscernible main beat (underlined), and the vowel deletion sign # in [εks#$^{\underline{\text{biʃʌn}}}$]. The substitution of [i] for [ɪ] in both [εks#$^{\underline{\text{biʃʌn}}}$] and in the two iterations of [$^{\text{it}}$] were also color-coded green. Conversely, red color-coding for "unintelligibile pronunciation" was employed for the misplaced stress symbol, the lack of word emphasis marker ◇ in "the-original-schedule," and the unreduced [^] and [ɪ] vowels in [$^{\text{ˎɔɹɪʤɜɪnʌl}}$].

Curiously, the incorrect stress placement and unreduced vowel combination[43] did not render the words [rimεm′bɜɹ] and [εks#$^{\underline{\text{biʃʌn}}}$] unintelligible as they did in [$^{\text{ˎɔɹɪʤɜɪnʌl}}$]. While it is, perhaps, difficult to explain as a general rule why intelligibility was compromised in the first two words listed above but not in the latter, there is an interpretative explanation which comes across in looking at the transcription. Speaker A's attack on the first syllable of [$^{\text{ˎɔɹɪʤɜɪnʌl}}$] (as though he was going to pronounce the word "orange")[44] coupled with the suspended pitch of evenly distributed syllables in his upper register, deflect the listener's attention away from understanding the word immediately so as to qualify for the definition of "unintelligible" that is operationalized in this study. In addition, there is a lack of distinction between important and unimportant words in "the-original-schedule," in contrast to the other two examples where the stressed word is clearly emphasized. The manner of articulation of the [l]

[43] These features often go hand in hand.

[44] Notably, the other 2 words start out comparatively well.

sound at the end of the word, which sounds swallowed, may also have contributed to a loss of intelligibility.[45]

Celce-Murcia et al. (1996) define stressed syllables as "those syllables within an utterance that are longer, louder, and higher in pitch" (p. 131). Given this definition, a listener would expect the word to be pronounced [ɔˈɹdʒɪnəl] or [ɔˈɹdʒənəl], with due emphasis on the second syllable.[46] In his articulation of [ˈɔɹɪdʒɪnʌl], however, Speaker A's syllables are, at least perceptively, equally long and equally high. (Loudness was not taken into account in the transcription and analysis because of the sound recording quality and speech editing procedures). Given the discrepancy between the two versions, it is not difficult to see why intelligibility might have been compromised.

From the above, it is clear that there are a number of pronunciation features at play, not least the listener's perception, that render a certain word intelligible or unintelligible. This passage was chosen because it exemplifies the conundrum of the beast known as "intelligibility" that applied linguists have been grappling with for so long. What is unclear is the very nature of intelligibility and the issue of how to demonstrate intelligibility and unintelligibility with empirical evidence. (For studies that addresses this second question methodically for one pronunciation feature, see Hahn, 2004; Field, 2005). Having demonstrated the complexities of the matter, future phonological examples in this thesis will henceforth be more clear-cut so as not to complicate an already complicated phenomenon.

Application of Morley's theoretical models.

Once the data-driven transcriptions and color-coding for intelligibility were complete, the analysis was informed by two of Morley's theoretical models, which were applied to the data in order to "double check" that features which have been deemed theoretically important by this central proponent of "new wave pronunciation" (Morley, 1994, p. 70) had been accounted for in the data analysis, with the goal of enhancing an understanding of intelligibility. The secondary intent was to take steps (albeit baby ones) towards the empirical validation of these models. To these ends, the features which were listed in the micro level column of Morley's *Dual focus: Speech production and speech*

[45] [l] was probably articulated as a retroflex rather than a lateral approximant.
[46] The superscript on the second syllable probably exaggerates this slight rise in pitch, however.

performance (1994, p. 75) were applied to the data in the manner of a checklist, in order to ensure that these "discrete points" had been taken into consideration in the phonological analysis. The only micro level feature that was not regarded in the analysis is the overall volume, since it was acknowledged that the speech recording quality and editing procedures may have altered the original volume of speech. As well, the speaker may simply have adjusted his /her habitual degree of loudness in the first place by virtue of the research condition of speaking into a microphone in a small room.

Morley's *Speech Intelligibility/ Communicability Index* (1994) was the second instrument to be applied to the data. While the strong communicative component of the index and an understanding of the relationship between intelligibility and accentedness across the two columns is beyond the scope of the study (see Literature Review), it was still possible to consider the extent to which the intelligibility column could be applied to the speech samples. While this will not be discussed in detail, suffice it to say that the process of trying to assign to each speaker an intelligibility score with this instrument was like trying to jam together two pieces of a puzzle that don't fit. The index would need to be refined, particularly in the mid-upper range of the scale, and the descriptors made more consistent and precise (and be limited to pronunciation) in order for it to be applicable to the insights derived in this study based on the speech samples of these 19 graduate students. In future studies, a data-driven rating scale might be the procedure to follow. (See Turner & Upshur, 2002).

Qualitative Analysis

Delineation of the process and procedure.

Open coding, which was loosely adapted from the breakdown of the analytic process in Strauss and Corbin (1998), was utilized in the qualitative analysis of the raters' comments with the final goal of generating categories that "emerge" from the data. The process was mostly cyclical rather than linear, but essentially consisted of a series of discernable steps. This involved: transcribing the data chronologically by rater, starting with Rater 1 (R1) and going through until Rater 18 (R18); grouping the data by Speaker with color-coding, from Speaker C (#C) to Speaker R (#R); merging the data for all raters and speakers; grouping the merged data according to some common thread; bolding the words that were deemed to be the most essential elements in the comments; generating categories for the grouped data using the raters' language; subdividing categories where appropriate; and finally, "imposing" my own language on the data to generate categories.[47]

The qualitative transcription and analysis was performed using *Microsoft Word.* The raters' original spelling and punctuation marks were retained (even in the examples provided in this thesis) in order to capture, as much as possible, their original language. It should be noted that the "Practice Speaker," Speaker D, was excluded from both qualitative and quantities analyses, since raters were told during the rating sessions that the "practice rating" was to be used for the sole purpose of familiarizing them with the data collection instrument and not for evaluating Speaker D's performance.

Two weeks after the initial qualitative analysis was complete, a check for consistency in the coding of the transcripts was carried out in accordance with procedures outlined by Johnson and Christensen (2004). The raw data were sifted through in random order, and the above steps were followed a second time without reference to what had been done earlier. Although the order of the comments within the categories was often different, the basic categories and subcategories generated were the same, indicating a high level of intracoder (or intrarater) reliability.

The qualitative dataset.

While it was expected that the raters would be diligent in filling out the quantitative part of the questionnaire, it was unexpected that the majority would choose to write comments as well. Seventeen out of 18 raters wrote comments, and it was noted that Rater 17, the only rater who didn't, was busily eating a pizza during the rating session. A third of the raters wrote comments for all 8 of the speakers that were evaluated in the rating session, and another third did so for 6 or 7 of the speakers. This is testament to the fact that the topic of this research study did resonate with the raters/ undergraduate students given the ITA context, and that they did have something to say about the speakers' speech. In fact, several of them mentioned to me that this study should look at professors' speech as well, intimating that their professors are sometimes difficult to understand.

[47] A more extensive account of the qualitative process and procedures can be obtained by e-mailing the author at talia.isaacs@elf.mcgill.ca

In addition to considerations of pronunciation and intelligibility that constituted the vast majority of the raters' comments, grammar, vocabulary, and meaning also figured into the comments, and were classified in their own distinct categories with various subcategories. While the sum total of the comments are insightful in shedding light on what undergraduate students notice in non-native speech, and, further, on which linguistic features might contribute to a broader definition of intelligibility that does not limit itself to pronunciation, the discussion in the Results chapter will be restricted to those comments that do center around pronunciation inasmuch as it relates to intelligibility.

Quantitative Analysis

All quantitative questionnaire data were analyzed using the statistical software package *SPSS 13.0*.

Justification of analyses conducted.

Appropriate, perhaps, to a descriptive study, descriptive statistics and frequencies essentially constitute the quantitative analysis. The statistical significance of ratings, however, was not calculated, since inherent in calculations of statistical significance is the assumption that the results of the study are generalizable to the entire population, and it was not felt that either the speaker or rater samples in this study are representative of their respective populations such that results can be generalized.[48] The first reason for this is that the study makes use of a convenience sample which is, by definition, a biased sample (Johnson & Christensen, 2004). That is, participants were not randomly selected based on probability measures. Rather, members of the two populations that were in some way "accessible" were invited to participate in the study, and those who consented became part of the participant trajectory. Secondly, the sample size of the speakers who were rated (*N*=8) and of the raters (*N*=18) is, without a doubt, too small for the sampling distribution of the mean to be calculated (Bachman, 2004) and for results to be generalized to the population (see Brown, 2001).

At this point, it should perhaps be emphasized that the goal of the study is not to generalize results to an entire population, but rather to use a few cases to elucidate our

[48] This decision was made in consultation with Yanhong Zhang of the *UNESCO Institute for Statistics*. Any oversight, however, is my own.

thoughts on intelligibility inasmuch as it relates to second language pronunciation. (For references on case study research see Yin, 2003 and Stake, 1995). The statistical analyses that were employed are a logical extension of this.

Results of the quantitative, qualitative, and phonological analyses will be looked at with respect to the research questions in the next chapter.

Chapter 4: Results and Discussion

In this chapter, evidence from the qualitative, quantitative, and phonological analyses that were conducted will directly address the two main research questions.

Research Question 1

The first question is as follows:

1. Is intelligibility "enough," that is, a sufficient goal and an adequate assessment criterion for evaluating proficiency in the pronunciation of non-native English speaking graduate students in the academic domain?

The short answer to this question is that yes, intelligibility is enough. Table 1 shows the intelligibility ratings that were assigned to each of the speakers in the Native Speaker Questionnaire/ Rating Scheme.[49] The percentages on the rating scale at the beginning of Section 2 were approximated based on where the "X" was marked for each of the speakers (except in cases where the raters themselves indicated the exact percentage that they had assigned), and these approximations were verified three times after they were entered in *SPSS* to ensure consistency in interpretation.

In looking at the data, one of the first things that jumps out is the variability in the ratings both within and across speakers. The high standard deviation is striking, amounting to 20.19 for all speakers, and the minimum and maximums values for each speaker seems to confirm that scores are all over the place. Nonetheless, we will see as the discussion develops that there are, in fact, discernable patterns in the data.

Table 1 shows that Speaker K is rated as the most intelligible. That is, the raters reportedly understood a larger percent of Speaker K's words than those of any other speaker. Speaker K received a mean rating of 95.22%, which is more than 10% higher than any other speaker and almost 20% above the group mean. Conversely, the speaker whose words are least intelligible to the raters is Speaker C, whose mean score of 46.94% is over 20% lower than the speaker with the second lowest rating and almost 30% below the group mean. In short, Speaker K and especially Speaker C stand out from the rest of the group in terms of being the most and least intelligible respectively.

[49] **In all tables pertaining to Research Question 1, which sort the data by speaker, the highest and lowest ranked speakers are bolded and the data tabulated for the group as a whole is bolded and underlined in the bottom row.**

Table 1

Mean Intelligibility Ratings

Speaker	Min. (%)	Max. (%)	Mean (%)	*SD*	Order
C	**8**	**86**	**46.94**	**18.51**	**8**
E	50	100	75.17	15.09	6
F	50	100	86.28	13.72	2
G	50	98	76.56	13.97	5
K	**75**	**100**	**95.22**	**8.54**	**1**
M	30	88	67.67	18.24	7
N	25	100	79.06	18.18	4
R	50	100	85.78	12.62	3
All	8	100	76.59	20.19	

Of the rest of the speakers, Speakers F and R, who are rated second and third most intelligible, differ by as little as half a percentage point from one another, Speakers G and E, who sit about 10% lower, are rated at fifth and sixth, and Speaker N, who is rated fourth at 79.06%, is sandwiched between the two groups (i.e., Speakers F and R and Speakers G and E). From there, there is around an 8% drop for speaker M, rated at 67.67%, followed by a plunge for Speaker C, the only speaker whose words are rated as less than 50% intelligible.

If we look at the scores for Speaker C, we see that ratings range from 8% to 86%. Similarly, ratings for Speaker N, which are a few percentage points above average, range from 25% to 100%. This variability in the ratings for each speaker attests to the fact that the raters are lay listeners with no background in pronunciation in addition to being untrained raters whose impressionistic scores have never been calibrated. That being said, the raters do seem to be more in agreement about Speaker K's intelligibility scores than about the scores of other speakers. This is shown by the standard deviation of 8.54 for Speaker K, which is more than 4.0 smaller than for any other speaker. Nonetheless, this standard deviation value is still very high by empirical research standards, which again drives home the point that these are not professionally trained *TSE* raters but rather undergraduate students who, having received brief instructions, are conferring ratings based on their initial impressions of the speech after just one listening.

The wide range and variability of scores for the speakers may also have been exacerbated by outliers, that is, extremely harsh or extremely easy raters whose ratings are not in line with the other ratings. These data are especially susceptible to outliers because of the small sample size (N=18 raters). In this dataset, all data that were deemed interpretable were retained for the analysis: no useable data were thrown out. Further qualitative and quantitative analysis will help show that there are some clear patterns in the data despite the small sample size.

Table 2 shows the frequency and means of comprehensibility ratings for each speaker. These data correspond to question 7 in Section 2 of the Questionnaire/ Rating Scheme, which features a 4-point Likert Scale. (See Appendix E). Figure 3 presents mean intelligibility and comprehensibility ratings side by side so that they can be readily compared. The speakers' mean intelligibility ratings (data from Table 2) are plotted on a scale on the left side of the page; the mean comprehensibility ratings are shown on the right side. Figure 4 graphs the raters' responses to questions 1 and 2 in Section 3, which asked them to list, in any order, a maximum of 2 speakers that stand out in terms of being the easiest and the most difficult to understand.

Table 2

Frequencies and Means of Comprehensibility Scores[50]

			Frequencies						
Speaker	*N*	Missing Data	Very Difficult	Difficult	Easy	Very Easy	Mean Score	*SD*	Order
C	18	-	**9**	**9**	-	-	**1.50**	**0.51**	**8**
E	15	3	-	7	7	1	2.60	0.63	5
F	18	-	-	2	15	1	2.94	0.42	2
G	18	-	1	10	6	1	2.39	0.70	6
K	17	**1**	-	**1**	**8**	**8**	**3.41**	**0.62**	**1**
M	18	-	1	11	6	-	2.28	0.58	7
N	18	-	1	4	9	4	2.89	0.83	4
R	16	2	-	4	8	4	3.00	0.73	3
All	**138**	**6**	**8.7**	**34.8**	**42.8**	**13.8**	**2.62**	**0.83**	

In looking at the data from Table 2, it is apparent that the general contour of the mean intelligibility scores and the mean comprehensibility scores are the same: Speaker C was rated as by far the least comprehensible of the speakers, considered by half of the raters to be "very difficult" to understand and receiving by far the lowest mean score, while Speaker K retained his position as the easiest speaker to understand, receiving the most "very easy" scores and the highest mean score. Evidence that these 2 speakers

[50] Any data that were considered ambiguous or uninterpretable appear in this table as missing data. This includes if the rater did not check off any boxes as well as if more than one box was checked off.

Figure 3

Mean Intelligibility Ratings versus Mean Comprehensibility Ratings[51]

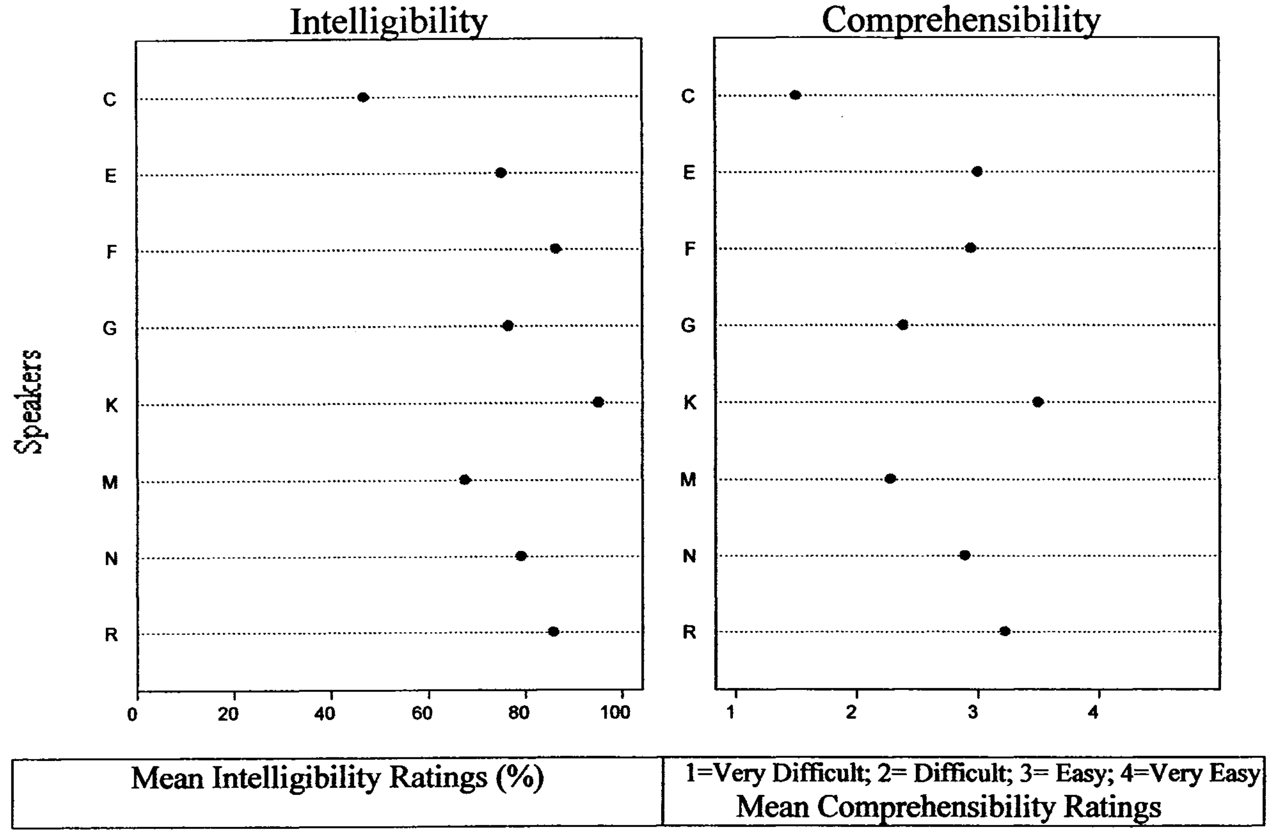
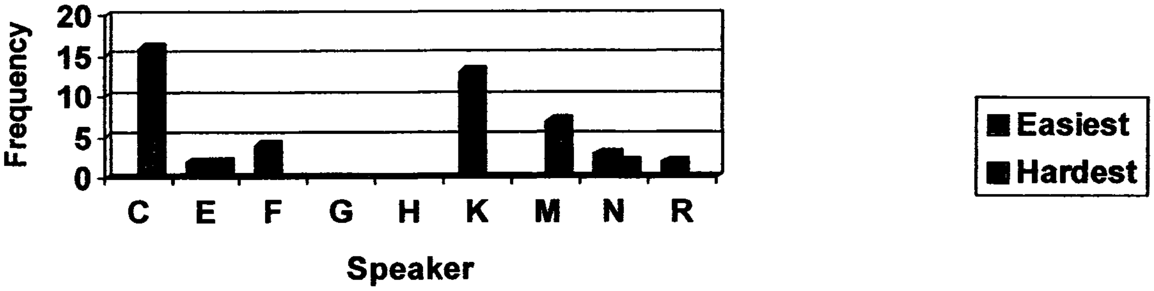


Figure 4

Easiest and Hardest Speakers to Understand Overall



[51] Yanhong Zhang created this figure for me using the statistical software package *Stata*, since this graphing function was difficult to perform using *Excel* and *SPSS*, the software that I had access to. Although the comprehensibility scale looks like it runs from 1 to 5, it really should only be from 1 to 4, where 1 is the smallest possible mean score and 4 the largest possible mean score.

stood out to the raters at the end of the session as well is plain from their respective "skyscraper" bar lines in Figure 4. Speaker M is also visible as difficult to understand overall, albeit to a lesser extent than Speaker C. This is consistent with the data in both Tables 1 and 2.

Resuming the comparison between the intelligibility and comprehensibility data in Tables 1 and 2, there was some minor shifting between the 2 "bookends" (Speakers C and K) but the general pattern was retained. Although Speaker R received a higher mean comprehensibility score than Speaker F, in assigning the ordinal rankings, it was felt that Speaker F should be ranked second, since the consensus among the vast majority of raters was that he was "easy" to understand. Conversely, for Speaker R, only half of the raters whose data were interpretable chose the "easy" designation, with each of the remaining quarters assigning "very easy" and "very difficult" respectively. This difference between the two speakers is reflected in the standard deviation values. Still, as with the intelligibility scores, Speakers F and R are really neck-in-neck for comprehensibility.

Another difference between Tables 1 and 2 that should be noted is the reversal of position between Speakers E and G. In Table 1, Speaker G, who was rated fifth, is closest to the group mean, and in Table 2, Speaker E, who climbed to fifth, is closest to the group mean. That Speaker E was thought by the raters to be somewhere in between "easy" and "difficult" is reflected in the comments of 2 individual raters. Rater 2 placed a checkmark between the "difficult" and "easy" boxes, writing "btwn the two," which, unfortunately was counted as "missing data;" Rater 2 put a checkmark next to "easy" and a second bracketed checkmark next to "difficult" with the note, "I would if I could." This was counted as an "easy" vote.

So although there seem to be some patterns between ratings for intelligibility, comprehensibility, and the speakers identified as the easiest and hardest to understand overall, what does this say about the primary research question of whether intelligibility is "enough" for graduate students in the academic domain? To properly address this, we will need to look at a little more quantitative data in addition to insights from the qualitative and phonological analyses.

The color-coding for unintelligibility for the whole set of 19 speakers (not just those included in the rating session) reveals that a much higher percentage of words were found to be intelligible in the phonological analysis than in the mean intelligibility scores assigned by the raters. Based on the 90 second duration of the speech samples, Speakers K, Q, and S were found to be 100% intelligible. That is to say, they produced no unintelligible words, (although for Speaker Q in particular, there were several coded instances of "unusual" pronunciation).[52] Moreover, speech samples from Speakers F, H, J, P, and R were all found to have less than three instances of unintelligibility. Speaker H, who had but one unintelligible word in her total of 160 words in the speech sample, could be considered 99.94% intelligible based on word count. Thus, almost half of the total sample of 19 speakers (i.e., the 8 speakers listed above) were over 99% intelligible. It is not necessarily the case, however, that the speakers with the most intelligible words were always the most comprehensible, as we will see later on in the chapter. Also to keep in mind is the fact that several variables, including a task effect, may have had an impact on intelligibility in *TSE* Item 12 that was phonologically analyzed, which was different from the two items that the raters rated.

Speaker C was found to have considerably more unintelligible words than any of the other speakers in the sample, with approximately 1 in 6 words being color-coded as unintelligible. That still makes her 83% intelligible, however, a far cry from the 46.94% that the raters assigned in the mean intelligibility ratings. To get a sense of the proportion of her words that are unintelligible as compared with the other speakers, we could say that Speakers' A, B, D, E, F, and G's combined total of 31 unintelligible words, amounting to 9 minutes of speech, is roughly equivalent to Speaker C's unintelligible words in 90 seconds of speech. Of course, this quantification of intelligibility, which is based on word count and dependent on rate of speech, must be taken with a grain of salt (as must all of the phonological analysis), since it is the product of one pronunciation researcher's analysis and interpretations of the speech samples after several listenings, a scenario which makes it difficult to determine what is and what is not immediately intelligible in accordance with the definition of the term that is employed in the study.

While a much higher proportion of intelligible words were detected in the phonological analysis than in the raters' intelligibility ratings, there was nonetheless

[52] We will remember that only those graduate students whose words were found to be 100% intelligible by the examiner were exempted from the second part of the *EPT Test* and ESL oral course at the University of Illinois Urbana-Champaign. (See Literature Review).

agreement between the raters and myself on the sometimes dire effects of unintelligibility. Rater 6, for instance, who assigned Speaker C an intelligibility score which was almost 15% above the mean, made the following comment:

> R6#C Easy to make out most words (60%), but difficult to make sense overall since 60% is all you understand!

This statement that Speaker C's words are easy enough to "make out" but that the overall meaning is difficult to comprehend conforms with my own "expert" opinion. Speaker C's speech *is*, without a doubt, difficult to understand. The proportion of 1 in 6 words being unintelligible is debilitating in terms of allowing the listener "to make sense overall." Although raters may have misjudged the proportion of unintelligible words in the same way that it is difficult to guess how many green jellybeans there are in a jar, they were definitely able to identify the most and least intelligible and comprehensible speakers. In fact, the ordinal ratings that were assigned based on intelligibility scores (Table 1) correspond almost exactly with my own intelligibility "rankings" of the 8 speakers based on the analysis and coding. Two minor differences of opinion are as follows: I found Speaker G's speech to be quite a bit less intelligible than Speaker E's, while the raters seemed to cast them in a similar light, evaluating Speaker G as slightly more intelligible than Speaker E in the intelligibility ratings, even though their position was reversed in the comprehensibility ratings. Second, Speaker E's response to *TSE* Item 12 was found to be slightly more intelligible than Speaker N's response, although Speaker N (ranked at a consistent fourth by the raters) spoke a lot more rapidly. Admittedly, Speakers G, E, and N, who place in the middle of the group, are difficult to distinguish in terms of intelligibility rankings, although their respective reasons for unintelligibility are drastically different as we will discover later on in the chapter. That the raters were so apt at rating them is quite telling.

There seems to be little consensus among the raters about Speaker N's intelligibility or comprehensibility. In fact, as we can see from the graph in Figure 3, some raters cited Speaker N as the easiest speaker to understand overall (*N*=3), while others found him to be the hardest to understand (*N*=2). In other words, for Speaker N, opinion was split. He was the only Speaker who was, in fact, rated as being both "very easy" and "very difficult" to understand in Table 2, although only one rater considered him to be "very difficult." Speaker N, in fact, poses a challenge to the very definition of "intelligibility" that is used in this study, since the definition doesn't make the distinction between important and unimportant words. In other words, there is no specification about which words must be intelligible – the criterion is just that the words must be immediately understood without having to guess at words.

This point also came up with Speaker F when considering the one instance of unintelligible speech that occurred in his entire response to *TSE* Item 12, which appears in the following passage:

Figure 5

How to Count Unintelligible Words

●
[aj‿dɪstɹɪb$^{\text{ˈju}}$tər‿ə‿æ$_{\text{-d}}$ _ ˈbifɔɹ bʌʔ aj‿ˈnidəd‿ty ʧejnʤ sʌm
i distributed a ad before but i needed to change some
●
ˈsʌm$_{\text{θɪŋ}}$ (1.) sow‿ »ʌm‿ənæ‿ʌm‿ənæ« tɛl‿jy_wʌɾ‿aj ʧ$_{\text{ejnʤd}}$ _]
something so um onna um onna tell you what I changed...

The displaced stress on [dɪstɹɪb$^{\text{ˈju}}$tər] was color-coded for unusual pronunciation but was not found to be unintelligible. What was unintelligible was the speaker's rapidly spoken [»ʌm‿ənæ‿ʌm‿ənæ«] (i.e., I'm gonna I'm gonna). This was counted as just one unintelligible word and not four since it sounds like one difficult-to-distinguish unit in the recording because of the linking and repetition. Although it is decidedly unintelligible and took several listenings to figure out what was being said, it is possible to see how any speaker, native or non-native, would have skimmed over these words, not enunciating them very clearly. This seems to be much less severe an instance of unintelligibility than an example of double substitution that came up in *TSE* Item 5 that the raters listened to, when Speaker C said "slipshot" instead of "snapshot" – a word which is much more fundamental to the meaning of the passage. The definition of intelligibility in this study, however, doesn't take such differences into account.

It is also likely that, although raters in all groups received the same "treatment," (i.e., teaching session, written and spoken instructions, and practice rating), they did not all internalize and interpret "intelligibility" in the same ways. Some raters who identified a given speaker as being 100% intelligible still rank ordered hindering features, whereas others made use of the "none" box in the same scenario. This suggests that some raters took the directions of only rank ordering those features that interfered with intelligibility more to heart than others, who identified features that were peripheral to intelligibility but still noticeable in the rank ordered boxes.

Fortunately, some qualitative comments, categorized under the heading ***A feature of the speaker's pronunciation that, while noticeable or irritating, does not affect overall intelligibility***, sheds some light on the matter. The comments are as follows:

> R1#F other than mispronunciation of certain words, fairly easy to understand.
> R7#G Easy to understand except for pauses & mumbling
> R11#E mumbles a bit… but not enough to be unclear.
> R13#K although I could understand everything I think he spoke way too slowly & overpronounced every indiv. word
> R15#K He overpronunced words & spoke slowly but it didn't really hinder comprehension
> R10#K understood all he said, but he spoke to slowly and had no intonation in his voice which made it quite obnoxious
> R18#F pronunciation is irritating but does not effect clarity of what is said.

These data lend additional credence to the idea that in certain cases, even though some of the raters found that the speaker was intelligible and rank ordered problematic features, it is not a given that all identified features did, in fact, hinder intelligibility or comprehensibility, even though this may have been the case. Many of the raters themselves were aware of this as the above comments show.

Let us now turn to the "TA question," or question 8 in Section 2 of the Questionnaire/ Rating Scheme. This is relevant to the research question since the ITA context is part of the "academic domain." Table 3 shows that none of the raters felt that Speaker C's pronunciation was sufficient for her to TA an undergraduate course, although 1 in 6 raters marked that they were not certain. Yet as is evident from the bold font, Speaker K's position at the number 1 ordinal ranking is usurped here by Speaker F, although the margin is small. Other than that, however, the order of speakers is exactly the same as in the comprehensibility ratings: Speakers R and N are very close to one another as are Speakers E and G.[53] Speaker M, the only speaker other than Speaker C

[53] In assigning the ordinal rankings, "no" was weighted more heavily than "not sure."

whose "no" votes outnumber her "yes" votes, figures in somewhere between Speakers G and C.

Table 3

Is the speaker's pronunciation adequate to TA?

		Frequency (Valid %)			
Speaker	*N*	Yes	No	Not sure	Order
C	**18**	**-**	**83.3**	**16.7**	**8**
E	18	50.0	16.7	33.3	5
F	**18**	**77.8**	**5.6**	**16.7**	**1**
G	18	50.0	27.8	22.2	6
K	17	76.5	11.8	11.8	2
M	18	27.8	38.9	33.3	7
N	18	61.1	16.7	22.2	4
R	18	61.1	5.6	33.3	3
All	**143**	**50.0**	**25.7**	**23.6**	

The reason for Speaker K's demotion from first to second position, although seemingly negligible given the numbers, can be explained by looking at quantitative and qualitative data with respect to raters' answers. Although Rater 16 had marked that Speaker K was 100% intelligible and "very easy" to understand, he checked the "no" box for the TA question, writing in the comments "way to slow to be a TA." This parallels comments from Raters 3 and 11, both of whom identified Speaker K as being 100% intelligible, "easy," and "very easy" to understand respectively and, in contrast to Rater 16, "yes" to the TA question.

> R11#K only a few problematic words. but would be quite annoying as a TA.
> R3#K Far too slow, as a TA, he would be quite boring.

Rater 18, who also marked Speaker K as 100% intelligible and had even guessed that Speaker K's first language was English (apparently not detecting any foreign accent), neglected to fill out the subsequent comprehensibility and TA questions, which figure into the "missing data" column in both Tables 2 and 3. In the comments section, however, he wrote, "sounds like someone with a good accent and no 'orating' skill at all."

In addressing the second research question, we will probe which features of Speaker K's speech might have prompted the above 4 raters to comment on his being "slow," "annoying," "boring," and with "no orating skill at all." Right now, suffice it to say that even though these raters found Speaker K to be 100% intelligible, they still seemed reticent about the idea of his TAing an undergraduate course. In spite of these negative comments regarding one of the "top performers" in this group of speakers, however (in terms of intelligibility and comprehensibility), we can see from the quantitative data that the vast majority of raters did not hesitate in answering "yes" to the TA question for either him or for most of the other speakers. Indeed, 6 out of 8 speakers received more "yes" responses than "no" responses to the TA question, although opinion for Speakers E and G was split between "yes" on one hand and "no" and "not sure" on the other.

This is still good news for the graduate students in this study (who are both current and prospective ITAs), the professors they work for, instructors of pronunciation-communication courses for graduate students or preparatory ITA courses, undergraduate students who are instructed by ITAs, and ultimately, the university at large. Rater 3, however, made the point to me informally at the end of the first rating session that it is far easier to understand speech when that is all that you need to pay attention to (and not worry about content) than when you are sitting in a science course and need to understand both speech (a vehicle for understanding content) and crucial subject-matter, often imbued with technical vocabulary that you (the student) are responsible for knowing and might even be tested on. This point is well taken as is Rater 5's comment about Speaker C, which speaks for itself: "I believe that without knowing before hand what topic this person was speaking of, it would have been significantly more difficult to understand what she was talking about." In all probability, topic familiarity does have a role to play in word-level intelligibility. The question that we must not lose sight of, as much of the literature keeps stressing, is intelligible to whom? (See for example Taylor, 1991).

This study does not purport to simulate a real ITA context. Indeed, given that the recorded prompts are listened to twice and played one after another, the whole scenario is somewhat artificial. The research situation does, however, simulate the assessment of non-native speakers' speech that takes place after the administration of either the *TSE* or the *SPEAK Test* at a university, although naturally with the *TSE* rating scale, raters are required to focus on much more than just pronunciation. (See Literature

Review). What the study does show is that with the help of a rating instrument that is geared toward pronunciation, native English speaking undergraduate raters are able to assess the most intelligible and the most unintelligible non-native English speaking graduate students after listening to only short speech samples. The same could apply for undergraduate students (both native and non-native) evaluating the speech of their non-native English speaking professors, for example. Future research needs to address this.

The data presented in this section support a definitive "yes" to the research question that intelligibility is an adequate assessment criterion in the academic domain, although with minor hesitations. Given that the raters are undergraduate students, the "academic domain" that is most relevant to this study is the ITA context. Speakers who perform well in intelligibility and comprehensibility ratings are also, by and large, deemed by raters to have adequate pronunciation to TA an undergraduate course. However, there is some indication that even for speakers who are found by select raters to be 100% intelligible, certain factors about their speech (which may or may not be related to pronunciation) make some raters wary of those speakers TAing undergraduate courses. Other data point to inconsistencies in individual rater behavior. For instance, both Raters 13 and 16 rated Speaker E as being 75% intelligible. Yet Rater 13 indicated that Speaker E is "difficult" to understand and doesn't have adequate pronunciation to TA, whereas Rater 16 marked that she is "easy" to understand and does have adequate pronunciation to TA. Despite the discrepancies in individual rater behavior, the general patterns for each of the speakers are clear.

A brief statement about whether a threshold level of intelligibility is detectable: the quantitative data show that there is a threshold level of intelligibility for graduate students in the academic domain and that Speaker C is decidedly below the threshold level. Where Speaker M places with respect to the threshold is less certain, as is the notion of whether there might be a second threshold level of intelligibility for non-native speakers who "excel" in intelligibility ratings. These are fruitful areas of exploration for future research.

Research Question 2

The second research question is stated:

2. Which features of pronunciation are most crucial for intelligibility?

Although no definitive answer can be offered to this question, which remains one of *the* central questions in pronunciation research (Field, 2005; Jenkins, 2000), qualitative, quantitative, and phonological analyses will work together to shed some light on the matter. Only after examining the speech patterns of various speakers as they relate to intelligibility, however, will a response to this question be attempted.

Table 4 shows two sets of data from the Rating Scheme/ Questionnaire. On the left hand side is a frequency count for the pronunciation features that the raters rank ordered as most hindering intelligibility after the second listening of each of the speech samples; on the right side is a look at the frequency of the most crucial features to intelligibility that were rank ordered at the end of the session.[54] Table 5 presents frequency counts of the descriptors in the small boxes that were checked off after the pronunciation features had been rank ordered in Section 2. (See Appendix E).

Table 4

Frequencies of Most Crucial Pronunciation Features for Intelligibility[55]

	Section 2					Section 3				
Pronunciation Feature	#1	#2	#3	Total	**Order**	#1	#2	#3	Total	**Order**
Speech Clarity	35	21	7	63	**2**	6	3	5	13	**2**
Rate of Speech	15	8	10	33	**5**	-	-	1	1	**6**
Pitch	4	16	13	33	**6**	-	2	-	2	**5**
Sentence Rhythm	13	14	15	42	**4**	1	3	5	9	**4**
Word Stress	21	27	16	64	**3**	2	6	6	14	**3**
Individual Sounds	43	28	15	86	**1**	9	4	1	14	**1**
None Identified	-	-	-	12	**-**					

We will remember that the raters were to identify only 1 of the 2 choices that were listed below each feature that they'd rank ordered. This is clear-cut for a feature like "rate of speech," where the speech is either "too fast" or "too slow," the two choices

[54] A maximum of 3 identified features were counted for the analysis in accordance with directions.
[55] The rank order of features in the "order" columns, which are bolded for each data set so that they can be visually easier to compare, was assigned based on #1 rankings rather than on total frequency.

Table 5

Frequency of Rank Ordered Features Identified in Section 2 [56]

Pronunciation Feature	Problem	Frequency	Order
Speech Clarity (59)	Overpronunces	11	12
	Mumbles	48	1
Rate of Speech (36)	Too Fast	12	10
	Too Slow	24	6
Pitch (32)	Pitch Change	15	9
	Monotone	17	8
Sentence Rhythm (48)	Distinguish Words	40	4
	Linking	8	11
Word Stress (64)	Stress Syllable	43	3
	Distinguish Syllables	21	7
Individual Sounds (83)	Substitute Sounds	44	2
	Delete/Add Sounds	39	5

representing opposite ends of the spectrum. However, for a feature like "individual sounds," the choices are not mutually exclusive: the speaker may well both "substitute sounds" and "add or delete sounds." Although the instructions directed raters to only identify the "most prominent choice" for the pronunciation features that had been rank ordered, several raters checked both boxes for "individual sounds," indicating on their forms that both options were a given speaker's most prominent problems. Rater 10, for example, wrote beside two checked boxes, "both apply substantial amounts → most hindering effects" for Speaker C. Since the rater's rationale was clearly explained, an exception was made and both options were counted in the frequency count.

From my phonological analysis, it was determined that Rater 10 was right: substituting sounds and adding or deleting sounds often do go hand-in-hand. I also found that segmental errors (or "individual sounds" as referred to in the questionnaire) resulted in unintelligibility in a substantial number of cases, and would also have ranked it overall at as the most crucial feature for intelligibility as the raters did in Sections 2 and 3. (See Table 4). While all of the 19 speakers except for Speaker S made segmental errors in their speech samples, however, it only negatively impacted intelligibility for a certain number of speakers. It also tended to hinder comprehensibility when the word in question had a high functional load.

[56] Frequency counts between the left side of Table 4 and Table 5 don't quite match up due to uninterpretable data.

Speaker I is among those speakers for whom segmental errors pose a substantial problem in the phonological analysis. In a brief post-data collection chat with her, Speaker I mentioned to me that, based on her observation of "puzzled looks" from native speakers, she had deduced that native speakers find it harder to understand single words that she utters than whole phrases. When I asked her for an example, she produced the word "fair" which she pronounced "fire." I got her to say a sentence with the word and still couldn't figure it out. It was only when she actually spelled it for me that I could understand what she was trying to say.

Figure 6 presents two examples of instances in Speaker I's speech where segmentals hinder the intelligibility of meaning-loaded words. Notably she has no problems with either word stress or sentence rhythm – segmental errors are the only real challenges to her intelligibility. In example 1, she says "police" instead of "place," a clear example of vowel epenthesis (i.e., the addition of an extra vowel sound), followed by "nightshow center" instead of "nature center," which she quickly corrects. In the second example she says the "diedline" instead of the "deadline" for entries, which was unintelligible to me when I was doing the orthographic transcriptions. Conversely, Speaker N's [dɛdlɛjn] (which sounds like "deadlane" with an Irish lilt), also a substitution error, was immediately understandable to me although color-coded as unusual.[57]

Figure 6

Loss of Intelligibility due to Substitution and Epenthesis

● ● ● ● ●
1. [də‿pə'lis‿ɪz‿'awso _ naj$^{\text{'tʃow}}$ 'sɛntəɹ nej'ʧəɹ _ 'nejʧəɹ 'sɛntəɹ]
the police is awso nightshow center nature nature center…

● ● ↑P4●
2. [ðə‿$^{\text{'daj}}$lajn/ 'akʧuwi‿ðə‿'dajdlajn fɔɹ/ 'ɛntɹis _ «ha-z bin‿ʧ$^{\text{ejndʒt}}$]
the diedline actually the diedline for entries has been changed…

[57] Due to the fact "nature center" and "deadline for entries" were derived from written prompts in *TSE* Item 12, the possibility that the confounding variable of reading ability (i.e., faulty reading strategies) may have been the cause of the unintelligibility rather than speech production cannot be discounted . (See Munro & Derwing, 1994). This was not a feature in the *TSE* items that the raters rated, however.

As with any well-constructed diagnostic passage, "Woodlands Nature Center Main Office" brought out many segmental problems for the speakers.[58] I agree with Anderson-Hsieh (1995) that getting the [u] - [ʊ] contrast necessary to pronounce "Woodlands" ['wʊdlændz] and not ['wudlændz] is far less important than pronouncing main [mejn] and not [mɛn] "men," for example, because of functional load. This, I would argue, is important for word-level intelligibility too.

Figure 7 presents a passage from the Practice Speaker, Speaker D, that contains a few intelligibility-threatening examples of sound deletion at the end of words. Sound deletion was also a problem for other Mandarin speakers in the sample such as Speakers B, E, and I, for Speakers C and G who are native speakers of Korean, and for Speaker M who is a native speaker of two Indonesian languages.

Figure 7

Loss of Intelligibility due to Deletion

[ændə‿mɑɹ# 'stuwəɹ# _ ʌ-m f faj#/ ɑɹts də'ɹɛktəɹs _ fɹʌ#- ðə-
and demar stuar um feigh arts directors from the
●!
mɛ _ mɛtrɔ'po^{liəntən} _ ʌ‿mju'^{zijəm}]
metropolintan uh museum (end of excerpt)

The raters were quite adept at pointing out the specific segmental problems of each of the speakers, although I would argue that most of the items listed below are distracting to the listener rather than actually crucial for intelligibility. Here are a few specific things that the raters cited:

R4#R Substitutes "th" sound for a "t" sound
R9#C says "r" sound instead of "l" sound
R12#G cam-er-a - overpronounces every syllable.
R12#M adds sounds & overpronunces - filem → film; misses → mrs.
R3#E Add 's' at the end, films.
R13#K also, he adds the "s" sound b/w words
R12#K He stresses the "s" sound in his words.

[58] The word "Woodlands," which proved to be especially problematic, was mispronounced in the following creative ways: ['uwrlnæz], ['owðɜɹlan], [wurlænz], ['wowldlan], ['udləndz], and ['wʊdlɑ#s].

R1#M or carries on the last syllable too long until next thought ie 'ssss'...
R18#M whistling 's' is piercing
R9#N cuts off "g" sound at the end of words

Speaker D performed well on *TSE* Items 5 and 7 where she could use her own vocabulary. Indeed, this is why she was chosen as the "Practice Speaker" for the rating sessions. However, the need to pronounce words that she was unsure of or unfamiliar with in Item 12 was difficult for her and did result in a loss of intelligibility for those words. Figure 8 shows how Speaker D got stuck on the word "exhibition," which was unintelligible. Task effect certainly played an essential part in Speaker D's performance.

Figure 8

Loss of Intelligibility due to Task Effect

● ● ● !
[wɪw bi- _ ɔn‿dʌ‿mej/ θəɹt'ijns _ 'mandej _ ænd‿ɛm _ ðə ɪksbɪk
will be on uh may thirteens monday and um the exhbic
● ● ●
_ ɛkʃɪbɪ _ 'bɪʃɪnd/ ɪs _ ɪts/ ɔn _ ɪts 'djɜɹɪŋ _ ə‿'pijɹi$^{\text{jəds}}$]
exhibi bitiond is its on its during a periods...

Now let's look more closely at what the raters identified for each of the speakers as the most prominent problem for intelligibility. Table 6 shows percent frequencies of the data from Table 5 sorted by speaker. In the data analysis, a maximum of three possible problems identified by each rater for each speaker were counted, in accordance with the instructions.

It so happens that the items that the raters identified correspond almost exactly with what was color-coded red and green (for intelligible and unusual pronunciation respectively) for each of the speakers in the phonological analysis. The numbers suggest that, among the problems that were cited for only one speaker, certain raters were sensitive to the fact that Speaker E does not link sounds, thus making her speech sound mechanical and disjointed, that Speaker F has an idiosyncratic downward pitch inflection at the end of his thought groups, that Speaker M does not use the reduced schwa vowel, thereby producing words in the phonological analysis such as ['fa$^{\text{jnɑl-}}$], ['pik$^{\text{ʃɜrs}}$],

Table 6

Frequency Distribution of "Most Prominent Problem" across Speakers[59]

Problem Cited by Raters	Speakers							
	C	E	F	G	K	M	N	R
Overpronunces	-	-	28%	-	-	24%	-	-
Mumbles	**72%**	22%	-	28%	-	41%	**67%**	29%
Too Fast	-	-	-	-	-	-	**50%**	-
Too Slow	-	-	-	22%	**67%**	-	-	-
Pitch Change	-	-	22%	-	-	-	-	-
Monotone	-	-	-	-	44%	-	-	-
Distinguish Words	33%	39%	44%	33%	-	28%	-	28%
Linking	-	22%	-	-	-	-	-	-
Stress Syllable	38%	-	44%	**56%**	-	**50%**	-	33%
Distinguish Syllables	-	-	-	-	-	28%	-	-
Substitute Sounds	**71%**	31%	22%	41%	-	41%	-	**50%**
Delete/Add Sounds	35%	**53%**	-	35%	-	47%	28%	-

['ʌntɪl], and ['rijzʌlts] (final; pictures; until; results), which often interferes with rhythm as well. As per the phonological analysis, Speaker M is the only one of the three whose intelligibility is at stake as a result of these "problems." Indeed, while the group of raters is excellent at characterizing the speech samples by writing telling comments, they often fall short of successfully making the distinction between features which are crucial to intelligibility and features which are annoying or noticeable but not absolutely critical. As it turns out, identifying features which are essential for intelligibility is an extremely difficult task. Pronunciation experts (including myself) have not yet managed to do so in a methodical, all-encompassing way, which was part of the impetus for this study. In light of this, I think these untrained raters did extremely well in identifying key features that are relevant to each of the speakers. The question is whether or not the identified problem area actually renders speech unintelligible.

Let us first turn to Speaker C since she was rated as being the least intelligible and comprehensible of the group. It was noted in my research log at the latter stages of my phonological analysis that "those (speakers) who have a consistent tempo are the most intelligible." Speaker C is certainly not among those speakers, as the following rater comments show:

[59] Percentages are indicated only when over 20% of the raters identified it as a problem and bolded when the frequency is 50% or above.

R6#C Irratic Speach.
R18#C rhythm very broken
R13#C she sort of stopped & started
R18#C slow to fast to stop to start.
R7#C Sometimes speaks quickly & incomprehensibly
R7#C She should slow down & speak more clearly
R1#C stutters, hesitates.
R10#C awkward pauses
R15#C lots of unsure pauses

The following example of Speaker C's speech from the phonological analysis in Figure 9 demonstrates the jarring stop-and-start effect of what I think the raters were trying to convey in their comments.

Figure 9

Loss of Intelligibility due to "Irratic Speach"

```
               !            !                                 -⊏
[»ṇ‿wɛn‿aj« (1.) wɛn‿aj wɛn‿ju (1.) hɑv tu kɔm in _ də/ dɪs
n when I when I when you have to come in da dis
                                  ⊐-      ●              ●
plejs ṇ so‿ju‿hæv tw _ kɔm _ ɔn 'fɹajde: _ m-ej‿ tᵋⁿ]
place n so you have to come on friday may ten...
```

Speaker C's abrupt sounding [»ṇ‿wɛn‿aj« (1.) wɛn‿aj] made it impossible to decipher, even after listening to the passage multiple times at a reduced speed. In other words, not only were these words not ***immediately*** intelligible to me, but they remain unintelligible to me. These are decidedly the most unintelligible few words in my entire 28 pages of phonological analysis and probably the most clear-cut instance of non-intelligible words in the analyzed data. The fast speed at which the slurred series of sounds is uttered certainly plays a large part in their unintelligibility.

Figure 9 also exemplifies that Speaker C's pauses occur at awkward junctures, (i.e., not at the end of clauses or thought groups). As I wrote in my research journal, "the staccatos make it seem as though she's been touched by fire. This messes up the rhythm." The lack of pitch variation, while not the cause of unintelligibility, certainly does not help the matter much. All of this gives the impression of poor articulation, which corresponds most closely with "mumbling" on the Rating Scheme/ Questionnaire which the raters identified. Indeed, both the qualitative and quantitative data show that raters were able to identify Speaker C's central problems, which are interrelated. In particular, the combination of Speaker C's not stressing the strong beat in the word and some sort of segmental error (usually substitution or deletion) is detrimental to intelligibility. This is evident in such words as ['ɪmpɔɹtʌn], [ɪzæmpu], and [fɹʌfɛs/'ɛr] (important; example; professor) which occur at different points in her answer to *TSE* Item 12, where the main beat is either displaced or not emphasized.

The other speaker for whom rhythm and pacing is an issue is Speaker N, whom, as we will recall from earlier on in the chapter, raters had a mixed opinion about in terms of intelligibility and comprehensibility ratings. The following comments, I believe, speak for themselves.

> R18#N pronunciate not too much of a problem...the train of thought being communicated is very broken.
>
> R18#N lots of words, repetitions, stutter → is most difficult part
>
> R11#N "ummm"s make sentences very choppy. (might just be nervous).
>
> R3#N Too fast and too much stuttering, too many "uh," "um," makes it quite confusing
>
> R1#N needs more time to think instead of blurting out fast words & stopping suddenly w/ ummm's.
>
> R15#N Too many "uhhh....'s" He pauses with an uhh.. and then speeds through the sentence until the next "uhhh"..
>
> R5#N The use of fillers such as 'uumm' or 'uhh' makes him unenjoyable & jerky to listen to.
>
> R13#N uses too many "ahhhs" & "umms"as if he's searching for the words & it's hard for him.
>
> R14#N Intermittent Gaps in vocabulary hinder proper rhythm.
>
> R1#N says 'umm' too much distracting
>
> R7#N Too many "um"s
>
> R2#N says like & um too much
>
> R7#N Too fast, too many ups & downs
>
> R12#N sometimes he talks to fast.
>
> R8#N spoke a bit too fast at times
>
> R1#N speaks too quickly & sinusoidally
>
> R10#N awkward pausing

Figure 10, an excerpt from Speaker N's answer to *TSE* Item 12, confirms what the raters were getting at in their comments. As we can see from the notation, "well she will not be turking part" is all uttered at the same pitch and at a very rapid pace, and then is followed by two prolonged "fillers," the duration of which was unfortunately not measured. In my research journal, an analogy was used which described Speaker N's speech as a motor of a stalled car which sputters a bit at first, madly accelerates at an uncontrollable pace as if catching on only to sputter again as before a few seconds later.

Figure 10

Loss of Intelligibility due to Bursts of Rapid Speech

●   -c
[«wɪl nɔt bij» ʌ- _ bij/ ʌ-‿»wə‿ʃij‿wɪl nɔt bij‿ˈtɜɹkɪŋ pɑɹt«‿jʌ-‿
will not be uh be uh well she will not be turking part yuh in
●   ●   ●   ●
ɪn ði‿ʌm-/ ɪn ði‿ʌm _ t ʌ-‿ɪn ðə _ bɔʀd‿ɔv‿ˈʤʌ$_{\text{ʤɪs}}$ _ æn‿ʌm]
the uh in the um t uh in the board of judges...

As hinted at earlier, Speaker N is an example of a speaker who is at times not intelligible but quite comprehensible (a distinction, of course, that the raters did not necessarily make). That is, although every single word that he utters in the rapid sequences is often not intelligible, my perception is that the message as a whole comes across well and is usually easy to understand. I would suspect that the perceived rapidity of his speech is exacerbated by the halting speech that often borders it on both sides. Speaker N would do best to tone down the fast parts so as to give his listener the illusion of evenness of tempo.

In the phonological analysis, Speaker J's speech is laboriously halting. While fast episodes such as Speaker N's are wholly absent from his speech, both speakers' speech samples are characterized by frequent repetition, occasional stuttering, and recurrent pausing, often in unnatural places. Although Speaker J has just one instance of unintelligibility in his passage, however, his thoughts are difficult for a listener to follow. In fact, I believe that Rater 18's astute comment for Speaker N, "pronunciate not too much of a problem...the train of thought being communicated is very broken" could as easily be applied to Speaker J as to Speaker N.

Speaker J's passage in Figure 11 exemplifies this. The pauses and repetitions in this figure underscore the fact that in the 14 lines of Speaker J's IPAed text, there are a total of 6 pauses which are over a second long and no less than 38 "ums."

Figure 11

Intelligible Speech which is Hard to Follow

• • • !• • •

[ʧejnʤ/ɪn plɑn‿ʌv‿ðə-‿ðə- ə'bawt‿ðə‿'fɔto/ fɔt fə'towgɹəfi kɔn't^{ɛst}
change in plan of the the about the phot photography contest

wɪʧ _wɪw‿bij‿ʌm (1.) ʌ _ w ˀ wɪʧ‿wɪw‿bij _ ʌ- _ də‿»də‿də«]
which wiw be um uh wh which will be uh da da da...[60]

Notably, Speaker J's pronunciation of the word "photography," which is provided in the *TSE* prompt, is sounded out with accurate word stress (although with a closed [o] rather than open [ɔ] sound on the second syllable). Since "photography" is on the ESL word stress hit list, I suspect that Speaker J might have encountered this word either while taking the graduate student pronunciation course offered for non-native English speaking graduate students at McGill, or in the explicit pronunciation training he had received. (For a summary of the speakers' backgrounds in pronunciation, see Appendix H). Following that, however, he stresses the wrong syllable for [kɔn't^{ɛst}], although 2-syllable nouns and verbs would undoubtedly have been covered at some point in his pronunciation instruction (e.g., the difference in pronunciation between CONtest-noun and conTEST-verb). Neither of these words affect Speaker J's intelligibility, however – his words are still immediately understandable when they are situated in context. The conclusion thus arrived at in examining Speaker J's speech is that it is not necessarily the case that slow, halting speech hinders intelligibility. Indeed, this may have little impact on the way the individual words are actually pronounced, although it may adversely affect comprehensibility. Conversely, insight from Speaker N's "jerky" speech suggests that while not all words are intelligible, the speech may still be by and large comprehensible.

The phonological examples that have been presented in reference to the second research question were chosen to coincide with qualitative and quantitative data from the rating sessions, where possible, in an attempt to unveil those features of pronunciation which are most crucial for intelligibility. As we have found, intelligibility can be

[60] Speaker J's next word, which unceremoniously got cut off in Figure 11 due to spacing, is "information."

compromised for different reasons, and is often the result of a combination of "problem areas" that interact together. Several examples in the phonological analysis have suggested, in accordance with previous research, that while adding and deleting sounds often hinder intelligibility, deleting sounds tends to have more of an inhibitory effect on intelligibility than adding sounds. (See Jenkins, 20000; Anderson-Hsieh, 1995). Substituting sounds can also be problematic for intelligibility, depending on the particular minimal pair in question and whether the word can be immediately deciphered in the context despite the slight mispronunciation.

As far as suprasegmentals are concerned, not emphasizing the primary stress in a multi-syllable word was a significant cause of unintelligibility, as was the need to accentuate important words in the sentence. The latter is like a road map for the listener. Indeed, the lack of emphasis on important words in the sentence, especially when coupled with another problem area like monotone speech or displaced stress, can make the speech more difficult to understand and perhaps even unintelligible. Pitch was not a problem for most speakers, with the exception of a few instances of wide fluctuation in pitch over a short period of time (i.e., one or two words). Not linking sounds between words can cause speech to seem disjointed, especially when pitch changes across words (as with Speaker E), but does not usually hinder intelligibility unless it interacts with other problematic pronunciation features. Speaker O, who has not yet figured into the discussion, sounds like she's making glottal stops word-initially due to an absence of linking, which causes frequent "outbursts of sound," as I described in my research journal. Again, this does not affect intelligibility, but it may be sufficiently distracting to the listener so as to cause unintelligibility when another problem area lurks in the background. Certain speakers may need to speed up or slow down their speech to enhance intelligibility, although uneven pacing may be tied to lapses in vocabulary, affective features (like nervousness), and other reasons which do not directly relate to pronunciation but can, nonetheless, be manifested in the way words are articulated and, thus, have an impact on intelligibility. Finally, pausing at the end of thought groups certainly makes the speech seem more "coherent." While not doing so, in itself, does not hinder intelligibility, it can threaten intelligibility if it is combined with other problematic pronunciation features.

In sum, the phonological analysis confirms that individual sounds, word stress, and speech clarity, the pronunciation features identified by raters as being most problematic in Table 5, can and often do hinder intelligibility in and of themselves,. Sentence rhythm, rate of speech, and pitch, on the other hand, tend to hinder intelligibility mostly when they operate in tandem with another pronunciation problem area.

Before closing the chapter, let us resume the query that we left off with in discussing the first research question, namely why some of the raters were put off by Speaker K's speech, expressing reticence to have him TA an undergraduate course in spite of his largely intelligible pronunciation, which was rated as substantially higher than all the other speakers. (See Table 1). In Table 6, raters identified Speaker K's speech as a combination of "too slow" and "monotone," although presumably these features had little effect on his intelligibility since the mean intelligibility rating that they assigned was over 95%. Figure 12 shows that Speaker K's speech is perfectly intelligible and lacks any unusual pronunciation.[61] Notable is the frequent use of the schwa vowel, in contrast to Speaker M's speech, for example, which lends his speech a certain smoothness and an English-like rhythm typical of a stress-timed language. (See Rogers, 2000). As is apparent in the rest of the speech sample, Speaker K also seems to be thinking consistently in clause units rather than word-by-word, leading him to pause at logical places.

Figure 12

Monotone but Highly Intelligible Speech

```
 ←                                    ●                            ●
[wəɹ 'gənæ‿ɹæp‿ʌp‿æt‿ðə‿ɹə'sεpʃən _ ɔn‿mej‿sεvən'tijnθ_
we're gonna wrap up at the reception on may seventeenth
               ●                                    ●    ●
fɹʌm fajv tə‿'sεvən _ æt‿ðə‿sejm 'gæləɹɪ‿ɹowm bij]
from five to seven at the same gallery room b...
```

Below are two thirds of the total raters' comments which were written in reference to Speaker K. Some of these comments have appeared at earlier parts of this

[61] Speaker K's use of the phrase "wrap up" is representative of the idiomatic expressions that are interspersed within his passage.

thesis. Nonetheless, they have been consolidated here to lend insight into the issue at hand.

> R10#K understood all he said, but he spoke to slowly and had no intonation in his voice which made it quite obnoxious
>
> R13#K although I could understand everything I think he spoke way too slowly & overpronounced every indiv. word
>
> R15#K He overpronunced words & spoke slowly but it didn't really hinder comprehension
>
> R12#K He speaks way too slow & monotone
>
> R7#K Speak faster & with more inflection
>
> R8#Ktoo slow and monotone
>
> R1#K would be very good if he sped up to create better flow
>
> R8#K speed up.
>
> R5#K I feel frustrated waiting for him to get on with what he is trying to say.
>
> R14#K Though slow, the choice of words is excellent and the effect is soothing.
>
> R3#K Far too slow, as a TA, he would be quite boring.
>
> R16#K way to slow to be T.A.
>
> R11#K only a few problematic words. but would be quite annoying as a TA.

These data confirm that Speaker K is "too slow and monotone," which is essentially an iteration of the information shown in Table 6. In addition, Raters 10 and 11 found his speech to be "obnoxious" and "annoying" respectively, although still mostly intelligible.[62] This notwithstanding, it is important not to lose sight of the fact that the vast majority of the raters (over 75%) still did see Speaker K as fit to TA. (See Table 3). By juxtaposing comments from Raters 3 and 5 on one hand and Rater 14 on the other, we can see that raters are not unanimously negative about Speaker K's slowness, although they all do seem to agree that he speaks slowly. I noticed that in my research journal, before having collected any data from the raters, I had written that Speaker K's "cold monotone makes him sound uninterested." Although this is peripheral to the narrow definition of intelligibility that is used in this study, it does have to do with sociolinguistics and attitudes towards speech. The ground is fertile for exploration in these areas.

[62] It should be remembered that the definition of intelligibility that is employed in this study excludes the notion of "irritability" (see Literature Review).

The concluding chapter will provide a brief summary of the research results, discuss the strengths and weaknesses of the study, and suggest future areas of inquiry related to the present topic.

Chapter 5: Conclusions and Implications

Summary

This study has attempted to address the two research questions by consulting data from quantitative, qualitative, and phonological analyses and arriving at an answer inductively. As defined for this study, intelligibility was found to be an appropriate goal and a suitable criterion for assessing proficiency in the pronunciation of non-native English speaking graduate students in the academic domain. The most intelligible speakers were also viewed by the raters as the most comprehensible and were also given the highest marks on the "TA question," although there is evidence that intelligibility is a necessary but not a sufficient condition to TA an undergraduate course. Furthermore, my own "expert" opinion as a pronunciation researcher who undertook extensive phonological analysis of the speech samples closely coincided with "untrained" raters' rankings of 8 of the speakers from most to least intelligible based on their mean ratings.

The consensus between myself and the raters is just as applicable to research question 2 as to question 1, where the approach has been to learn which pronunciation features are most important for intelligibility by identifying first what is unintelligible and what causes unintelligible speech and then proceeding to infer what might be most crucial for intelligibility. Of the features listed in Section 2 of the Questionnaire/ Rating Scheme (Appendix E), both myself and the raters identified "individual sounds," "speech clarity," and "word stress" as contributing the most to intelligibility. (See Table 4). Moreover, insights from the phonological analysis suggest that other features, such as "sentence rhythm," "rate of speech," and "pitch" tend to act in combination with other pronunciation problem areas to yield instances of unintelligibility.

This study contributes to our understanding of intelligibility in pronunciation, an area about which much has been said but that, in reality, we know little about. The results of the study have resonance in the areas of pronunciation assessment, pedagogy, and curriculum design. This concluding chapter will outline some of the methodological strengths and weaknesses of the study as well as propose suggestions for future research.

Strengths

"Intelligibility" is problematic in the pronunciation literature inasmuch as there is no field-wide consensus on to how to define or measure it. In this study, the theoretical and operational definitions of "intelligibility" were adapted from the *English Placement Test* (*EPT*), in use at the University of Illinois at Urbana-Champaign, and imported into the research context of this study at McGill University. The original definition of intelligibility from the *EPT* and the modified version employed in the present study have not, to my knowledge, been used in any other research context. The study thus presents a novel way of defining and measuring intelligibility for research purposes. The word-level focus that is entailed in its theoretical definition facilitates the construct's quantification in its operational definition, since the number of words that are understood is more obviously countable than a definition which incorporates anything to do with meaning or understanding the speaker's intended message. At the same time, the operational definition captures the raters' impressionistic ratings of the speech samples by getting them to plot points on a rating scale after just one listening. This essentially represents the raters' first impressions of a given speaker's intelligibility.

The Questionnaire/ Rating Scheme, which was specifically designed with the undergraduate raters in mind, successfully elicited meaningful data which address the central research questions. The instrument helped focus the raters' attention on intelligibility in the first listening and on identifying those pronunciation features that may have hindered intelligibility in the second listening - two tasks which correspond with the two research questions. The sheer volume of comments that the raters produced, several of which were "bang on" about some aspect of the speakers' speech patterns, speaks to the likelihood that the raters *did* have something they felt was worthwhile to say about the speakers' speech. It is also possible that the raters, who, as paying "customers" for their undergraduate education in the university reality are directly impacted if their ITA is unintelligible to them, felt personally involved in the study.

From the speakers' end of things, given that 13 of the 19 have taught in either EFL or, more crucially, ESL contexts, and that 11 cite teaching English as a professional activity that they envision after they graduate (see Appendix H), they require a certain degree of intelligibility as English language educators. This is in addition to the intelligibility required of them right now to carry out their academic tasks as graduate students, and for 2 of the speakers included in the rating session to carry out current instructional duties as ITAs. While this study focuses mostly on the ITA context, it also does tap into a population of speakers for whom intelligible pronunciation is important in other contexts as well. (See Morley, 1991). In short, the two groups of participants, who are strong stakeholders as far as intelligibility is concerned, strengthen the practical value of the study's implications. The use of the *TSE* as the speech elicitation instrument further enhances the relevance of the study to issues of graduate student admissions and ITA screening.

While the results are not generalizable beyond the study due to small sample size and a lack of random sampling, they do show some definitive patterns in the particular "cases" of the 8 speakers that were featured in the rating sessions. A certain degree of reliability in the quantitative data is evidenced by the fact that the relative order of the speakers' means in the intelligibility and comprehensibility scores, constructs which were judged to be theoretically similar (but not identical), match up.

Perhaps the greatest strength of the study lies in the interplay between the qualitative, quantitative, and phonological analyses (i.e., mixed design), each of which tells a different side of the story in the exploratory nature of the study. Considering all three is an essential part of trying to answer the quasi-philosophical research questions, which touch on the nature of intelligibility and what it is comprised of. The transcription system that was developed for the suprasegmentals shows how phonological transcription and coding systems can be data-driven, on the path towards finding a more efficient way to notate the "musical aspects of pronunciation" (Gilbert, 1994, p. 38). Furthermore, my attempt to establish intrarater reliability in the qualitative analysis is an exercise which did entail some conscious justification about the way the data were grouped. The overall congruence between the untrained raters' quantitative and qualitative data on one hand, and my own opinion after having gone through detailed phonological analysis on the other, lends strength to the findings.

The recognition of the need to empirically validate intelligibility models is crucial in order for pronunciation research to go beyond its current reliance on anecdotal evidence. While the intention behind applying Morley's *Intelligibility/ Communicability Index* to the data was noble, it remained nothing more than an intention and could not feasibly be implemented in this study. Future studies should strive for more than just talk in this regard. The idea of coming up with a data-driven intelligibility scale rather than trying to fit data into a pre-ordained system also has merit.

Weaknesses

The problem of defining and measuring intelligibility that pervades research on intelligibility also permeates this study. It will be remembered from the Literature Review that the use of a rating scale to measure intelligibility impressionistically assumes that intelligibility is incremental rather than an all-or-nothing phenomenon. Although the impressionistic ratings sought out in this study capture raters' initial impressions of the speech samples, the degree of subjectivity that is entailed exceeds measures of intelligibility employed in various studies by Derwing and Munro, for example, where the listeners actually write down what they hear from speech samples and the number of accurately deciphered words are coded and quantified based on a system established by the researchers. The way intelligibility is defined and operationalized in this study does not fit into Derwing and Munro's more restrictive definition.

There are two major arguments which challenge the way intelligibility was measured in this study. On one hand, it is perhaps not objective enough to be reliable. The considerable variability among raters in assigning intelligibility scores (as represented by the standard deviations) may not have been as high if a more objective way of measuring the construct (e.g., a dictation) had been employed instead. Further, it is also almost certain that raters' differential conceptions of what was being measured played into their judgments, even though raters were briefed about the goal of the study and every attempt was made to define intelligibility for them in a clear and accessible way.

On the other hand, it might be argued that in the real world of communication, it is not important to understand every single word that is uttered (which rarely happens with either native or non-native speakers anyway), but rather to understand the overall meaning or get the gist of the message. This point throws into question the real-world value of a definition of intelligibility that scrutinizes individual words, and would lean towards a concept of the term that embodies the broader goal of communication. As it happens, the definition of intelligibility in this study represents a sort of happy medium between Kenworthy's (1987) notion of "comfortable intelligibility" and Derwing et al.'s

(1998) stricter, more linguistically-based interpretation of the term. (See Literature Review).

Another problem with the way that intelligibility is defined in this study relates to the relationship between its theoretical and operational definitions. The requirement that words be immediately understandable to the listener without needing additional context is more compatible with the procedures in the rating sessions, where raters had to make intelligibility judgments after just one listening, than with the procedures in the phonological analysis, where intelligibility judgments were made after the multiple listenings necessary to perform a multi-layered phonological analysis. Indeed, gauging which exact words were immediately understandable in doing the color-coding near the end of the process was no easy task, since it was almost impossible to listen to the speech samples with fresh, unbiased ears. Although the original orthographic transcriptions gave clues as to which words had not been understood originally, even this notation process had required several listenings at reduced speed and thus was not fool-proof. In short, the notion of being able to understand a word immediately was more feasible for raters in the rating session than for myself, who had been submerged in the phonological analysis for an extended period of time, although the raters' assessments were more impressionistic and much less precise. Related to this, although efforts were made to provide self-checks on the data that I analyzed, my multifarious role as instrument constructor, data collector, data analyzer, and phonology "expert" may have been problematic in terms of "balance-of-power." Having another more objective pronunciation researcher involved in the analyses would have allowed interrater reliability in addition to the intrarater reliability employed in this study. This was not feasible within the scope of this study.

The use of the *TSE* as the speech sample elicitation instrument assumes that the *TSE* tasks are sufficient for the purpose of assessing proficiency in pronunciation and intelligibility. Yet no steps were taken in this study to ensure that this would be the case.

In selecting and preparing the speech samples for the rating sessions, care was taken to randomize the speakers in each of the rating sessions after the trial run with the same Practice Speaker. However, the *TSE* items themselves were always presented in the same order for each speaker in each of the rating sessions, namely picture sequence (Item 5) followed by expressing an opinion (Item 7). While it is unclear whether the order of these items affected the results, in half of the rating sessions the *TSE* items could have been presented in reverse order to ensure that there was no ordering effect.

There are a series of additional practical constraints on the study. The small sample sizes for speakers (*N*=19) and raters (*N*=18); the inclusion of a small subset of speakers in the rating session (*N*=8); the small number of *TSE* items which were included in the rating session (*N*=2); the small number of *TSE* items which were transcribed (*N*=4) and analyzed phonologically (*N*=1), etc. These are what are known as research realities.

Implications and Future Research

We will recall from the Literature Review that, in addition to using the *Test of Spoken English* and the *SPEAK* test to screen prospective ITAs, the University of Illinois at Urbana-Champaign has developed its own spoken assessment instrument, the *EPT*, for non-native English speaking graduate students who do not meet the institution- and/ or department-set cutoff score on the *TOEFL*. McGill University, in contrast, has not required any form of spoken assessment for its non-native English speaking graduate students as part of its admissions requirements as of the 2005-2006 school year, nor does it screen ITAs for speaking proficiency.

The results of this study suggest that it may be advantageous to assess not only the spoken ability but also the pronunciation of non-native English speaking graduate students. While the next generation *TOEFL* test, which is likely to be integrated into university admission requirements in the coming years, will incorporate a new speaking component which features "intelligibility" in its independent and integrated scoring rubrics (Educational Testing Service, 2005), this does not reduce the need for an assessment instrument which focuses specifically on pronunciation (and not just pronunciation as merely one of several components). Such an instrument would establish a certain proficiency standard for the pronunciation of non-native English speaking graduate students and/ or potential ITAs, and could conceivably be used for screening, diagnostic, or placement purposes depending on its intended use.

As the results of the study show, intelligibility is well-placed to be a useful assessment criterion for such an instrument. Empirically determining a threshold level of intelligibility, or the minimum intelligibility necessary for non-native English speaking graduate students to carry out their academic tasks (i.e., the lowest common denominator of intelligibility) would be a step in that direction. In this study, the phonological, quantitative, and qualitative analyses revealed that Speaker C is below the intelligibility threshold, whatever that threshold may be. Further research would need to define the threshold with more precision and explore whether there might also be a second threshold as Morley's *Intelligibility/ Communicability Index* intimates (1994) with her two "communicative thresholds." In order to evaluate this theoretical model, the link between "intelligibility," "accentedness," and "communicability" would need to be made more explicit.

As for pronunciation test construction, Saif (2002) reports on a needs-assessment approach for ITAs at the University of Victoria which utilizes Bachman and Palmer's (1996) test development framework to foster the link between the purpose of the test, the context in which it will be used, the Target Language Use Domain, etc. (Saif's study did not focus on pronunciation, however). This methodical approach to test development might be especially useful as a starting point at institutions which lack strong test development traditions upon which to draw. This approach could readily be geared towards developing an authentic instrument.

In her "Notes on Speech Evaluation" which accompany her *Speech Intelligibility/ Communicability Index*, Morley suggests, "try to listen to the speech sample as if you were an untrained language listener. Err on the conservative side with consideration of the 'lay' listeners whom the student will meet" (p. 77, 1994). Of course, depending on what purpose the index is being used for, this consideration may or may not be important. But inasmuch as the ITA context is concerned, why not incorporate undergraduate student "lay listeners" in the screening session?

Echoing Morley's "'lay' listener" suggestion, Porter and Weir (1997) argue that "at least part of the validation of criteria for assessing proficiency in pronunciation must be the gathering of information on what ordinary (non-linguist, non-applied-linguist, non-language-teacher) language users react to in the pronunciation of learners, and what the nature of the reaction is" (p. 25). In an attempt to establish an appropriate assessment criterion for proficiency in the pronunciation of graduate students and especially ITAs, again, would it be advantageous to incorporate undergraduate student "lay listeners" in the process rather than, as Morley (1994) suggests, just pretending to have them on board or trying to see through their lenses?

The present study has shown that, far from being "naïve listeners," untrained undergraduate raters have the capacity to be extremely astute, focused listeners when it comes to pronunciation. It is probable that the Questionnaire/ Rating Scheme helped the raters direct their attention to those aspects of speech that were most pertinent to the study. Their off-the-cuff impressionistic ratings and qualitative comments correspond very closely with my own "expert" assessments that were arrived at after many hours of being deeply engaged in phonological analysis. It must be acknowledged, however, that the raters who participated in this study were self-selected volunteers, and that there is no guarantee that all undergraduate students from the same population would have yielded the same quality of data. An undergraduate student voice is, nonetheless, instrumental in determining an acceptable pronunciation assessment standard. Not only are undergraduate raters the lay listeners that Morley (1994) and Porter and Weir (1997) call for, (unless, of course, they are linguistics or TESL majors), but they are also one of the main stakeholders in the ITA context since they are directly affected by their ability or inability to understand an ITA.

Future studies could cast their focus on which pronunciation features undergraduate students tend to identify when they listen to the speech of non-native English speaking graduate students or professors and how their attitudes shape their perceptions of intelligibility. Examining intelligibility as it relates to speech production and speech perception is another avenue to explore which is, perhaps, essential to understanding a concept which embodies the notions of utterance sender and utterance receiver in its scope. It would also be intriguing to compare the ratings of native and non-native raters coupled with, for example, surveying the relationship between accent familiarity and intelligibility ratings.

Another point of exploration which could be examined using the data that were elicited from the rating sessions is the issue of rater behaviour. This fits into the paradigm of "rater research" that has become a hot topic in second language assessment. In one line of inquiry, for example, Kim (2005, April), Turner & Upshur (1999), and Upshur & Turner, (2002) used a facet approach (IRT item response theory) to look at issues of rater severity. Some raters may be more consistently lax in their ratings and some more severe. Some ratings may show certain raters to be either positively or negatively biased towards a particular speaker, and some speakers tend to divide rater opinion more than other speakers. As well, some raters appear to be particularly sensitive to certain features of pronunciation while others show no indication of having noticed those very features. Raters' "habits" in filling out the questionnaires (e.g., whether they tend to identify the same features for all speakers across the board, vary their choices according to speaker, choose arbitrarily, etc.) could lend some insight into the different ways in which the raters interpreted intelligibility and the different strategies that they used in identifying their answers. Quantitative and qualitative data for each rater would ideally operate in tandem to provide essential clues about rater attitudes and behaviour.

By identifying those pronunciation features which are most crucial for intelligibility, the study has some important implications for ITA training and assessment and for the design of graduate student pronunciation courses. Such courses could ensure "more bang for the buck" if they targeted those features of pronunciation which are most critical for intelligibility and, indeed, combinations of features that are most detrimental to intelligibility. This would be a step up from current pedagogical practices which are largely based on theory or anecdotal rather than empirical evidence (Derwing et al., 1997), and could function to provide more focused pronunciation instruction. This, in turn, would benefit non-native English speaking graduate students, ITAs, the professors that employ and mentor them, the undergraduate students whom the ITAs instruct, and ultimately, the university at large.

This thesis will act as a springboard for my future doctoral work, which will make the link between pronunciation, diagnostic testing, and implications for a multi-dimensional curriculum more explicit.

Since this is an exploratory study, none of the findings are conclusive and the quantitative, qualitative, and phonological analyses all entail some degree of interpretation. What is clear is that intelligibility and unintelligibility are complex constructs, the meaning of which depends on how they are defined and measured. As a phenomenon which continues to be shrouded in obscurity, it is likely that intelligibility will continue to fascinate applied linguists and be a focus of research well into the next generation.

References

Abercrombie, D. (1956). *Problems and principles in language study*. London: Longman.

Anderson-Hsieh, J. (1995). Pronunciation factors affecting intelligibility in speakers of English as a Foreign Language. *Speak Out, 18*, 17-19.

Anderson-Hsieh, J., Johnson, R. & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42*, 529-555.

Atechi, S.N. (2004). *The intelligibility of native and non-native English speech: A comparative analysis of Cameroon English and American and British English.* (Doctoral dissertation, Chemnitz University of Technology). Retrieved December 6, 2005, from http://archiv.tu-chemnitz.de/pub/2004/0088

Bachman, L.F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

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

Bauer, G. (1996). Addressing special considerations when working with international teaching assistants. In J.D. Nyquist & D.H. Wulff (Eds.), *Working effectively with graduate assistants* (pp. 84-104). Thousand Oaks, CA: SAGE Publications.

Bauer, G. & Tanner, M. (Eds.). (1994). *Current approaches to international TA preparation in higher education: A collection of program descriptions*. Seattle, WA: Washington University.

Briggs, S.L. (1994). Using performance assessment methods. In C.G. Madden & C.L Myers (Eds.), *Discourse and performance of international teaching assistants* (pp. 63-80). Alexandria, VA: Teachers of English to Speakers of Other Languages.

Brown, A. (1989). Some thoughts on intelligibility. *The English Teacher, 18*. Retrieved August 25, 2005, from http://www.melta.org.my/ET/1989/main4.html

Brown, J.D. (2001). *Using surveys in language programs*. Cambridge: Cambridge University Press.

Caldwell, K. & Samuel, C. (2001). Reviews/ Critiques [Test of Spoken English (TSE)]. *Canadian Modern Language Journal, 58*(2), 319-326.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1,* 1-57.

Celce-Murcia, M., Brinton, D. & Goodwin, J. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages.* Cambridge: Cambridge University Press.

Cook, V. (Ed.). (2002). *Portraits of the L2 user.* Clevedon, UK: Multilingual Matters.

Crystal, D. (1997). *English as a global language.* Cambridge: Cambridge University Press.

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

Dalton, C. & Seidlhofer, B. (1994). *Pronunciation.* Oxford: Oxford University Press.

Dauer, R.M (1993). *Accurate English.* Englewood Cliffs, NJ: Prentice Hall Regents.

Derwing, T.M. & Munro, M.J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 20,* 1-16.

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

Derwing, T.M., Munro, M.J. & Wiebe, G.E. (1997). Pronunciation instruction for "fossilized" learners: Can it help? *Applied Language Learning, 8*(2), 217-235.

Educational Testing Service. (2005). *TOEFL iBT Tips.* Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2001). *TSE and SPEAK score user guide.* Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1995). *Test of Spoken English: Standard-setting manual.* Princeton, NJ: Educational Testing Service.

Eisenhower, K. (2002). *American attitudes toward accented English.* Unpublished M.A. thesis, McGill University, Montreal.

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

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

Gass, S. & Varonis, E.M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34*(1), 65-89.

Gatbonton, E., Trofimovich, P. & Magid, M. (2005). Learners' ethnic group affiliation and L2 pronunciation accuracy: A sociolinguistic investigation. *TESOL Quarterly, 39*(3), 489-512.

Gilbert, J.B. (1994). *Intonation: A navigation guide for the listener.* In J. Morley (Ed.), *Pronunciation pedagogy and theory* (pp. 34-38). Alexandria, VA: Teachers of English to Speakers of Other Languages.

Gimson, A.C. (1980). *An introduction to the pronunciation of English* (3rd ed.). London: Edward Arnold.

Goodwin, J., Brinton, D. & Celce-Murcia, M. (1994). Pronunciation assessment in the ESL/EFL curriculum. In J. Morley (Ed.), *Pronunciation pedagogy and theory* (pp. 5-16). Alexandria, VA: Teachers of English to Speakers of Other Languages.

Gordon, R.G. (Ed.). (2005). *Ethnologue, Languages of the World* (15th ed.) [Electronic version]. Retrieved December 6, 2005, from http://www.ethnologue.com

Graham, J.G. & Picklo, A.R. (1994, March). *Increasing relevancy in a speaking course for graduate students.* Paper presented at the Annual Meeting of Teachers of English to Speakers of Other Languages, Seattle, WA.

Grove, C. (1994). International TAs' oral language proficiency. In G. Bauer & M. Tanner (Eds.), *Current approaches to international TA preparation in higher education: A collection of program descriptions* (pp. 99-110). Seattle, WA: Washington University.

Hahn, L.D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*(2), 201-223.

Hahn, L.D. & Dickerson, W.B. (1999). *Speechcraft: Discourse pronunciation for advanced learners.* Ann Arbor, MI: University of Michigan Press.

Han, Z. (2004). *Fossilization in adult second language acquisition.* Clevedon, UK: Multilingual Matters.

Harper, D. (2001). *Online etymology dictionary.* Retrieved December 6, 2005, from http://www.etymonline.com/index.php?l=i&p=9

Hoekje, B. & Williams, J. (1994). Communicative competence as a theoretical framework for ITA education. In C.G. Madden & C.L Myers (Eds.), *Discourse and performance of international teaching assistants* (pp. 11-26). Alexandria, VA: Teachers of English to Speakers of Other Languages.

International Phonetic Association. (1999). *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press.

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

Johncock, P. (1991). International teaching assistants tests and testing policies at US Universities. *College & University, Vol. LXVI*, 129-137. Washington, DC: AACRAO.

Johnson, B. & Christensen, L. (2004). *Educational research: Quantitative, qualitative and mixed approaches* (2nd ed.). Boston: Pearson Education.

Kelly, L.G. (1969). *25 centuries of language teaching*. Rowley, MA: Newbury House.

Kenworthy, J. (1987). *Teaching English pronunciation*. London: Longman.

Kim, Y. (2005, April). *Are they really different? An investigation into native and non-native teachers' judgments of oral English performance*. Presented at the Research Exchange Forum, McGill University, Montreal.

Koren, S. (1995). Foreign language pronunciation testing: A new approach. *System, 23*(3), 387-400.

Lindley, M. & Campbell, M. (2005). Interval. *Grove Music Online*. Retrieved December 7, 2005, from www.grovemusic.com

Lado, R. (1961). *Language testing*. New York: McGraw-Hill Book Company.

Ludwig, J. (1982). Native speaker judgments of second language learners' efforts at communication: A review. *Modern Language Journal, 66*, 274-283.

McLaughlin, B. (1978). *Second-language acquisition in children*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mendenhall, M.E. (1996). The foreign teaching assistant as expatriate manager. In D. Landis & R.S. Bhagat (Eds.), *Handbook of intercultural training* (2nd ed. pp. 231-243). Thousand Oaks, CA: SAGE Publications.

Merriam-Webster. (2005). *Merriam-Webster Online Dictionary*. Retrieved December 6, 2005, from http://www.m-w.com

Morley, J. (1994). A multidimensional curriculum design for speech-pronunciation instruction. In J. Morley (Ed.), *Pronunciation pedagogy and theory* (pp. 64-91). Alexandria, VA: Teachers of English to Speakers of Other Languages.

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

Munro, M.J. & Derwing, T.M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 41*(1), 73-97.

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

Nelson, C.L. (1992). My language, your culture: Whose communicative competence? In B.B. Kachru (Ed.), *The Other Tongue: English across cultures* (pp. 327-339). Urbana, IL: University of Illinois Press.

National Institute on Deafness and Other Communication Disorders. (2005). *Autism and communication*. Retrieved December 6, 2005, from http://www.nidcd.nih.gov/order/index.asp

Nunberg, G. (2000). Usage in the American Heritage Dictionary. *The American Heritage Dictionary* [internet edition]. Retrieved December 6, 2005, from http://www.bartleby.com/61/7.html

Oppenheim, N. (1998, March). *Undergraduates' assessment of international teaching assistants' communicative competence*. Paper presented at the Annual Meeting of Teachers of English to Speakers of Other Languages, Seattle, WA.

Oppenheim, N. (1997, March). *How international teaching assistant programs can prevent lawsuits*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Porter, D. & Weir, C. (1997). Valid criteria for the assessment of proficiency in pronunciation. *Speak Out, 20*, 22-28

Prator, C.H. (1967). *Manual of American English pronunciation* (Rev. ed.). New York: Holt, Rinehard and Winston.

Pullum, G.K. & Ladusaw, W.A. (1996). *Phonetic symbol guide* (2nd ed.). Chicago: University of Chicago Press.

Raux, A. & Kawahara, T. (2002). *Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning.* Retrieved December 6, 2005, from http://www.cs.cmu.edu/~antoine/papers/icslp2002a.pdf

Rogers, H. (2000). *The sounds of language*. Harlow, UK: Pearson Education.

Rogerson, P. & Gilbert, J.B. (1990). *Speaking clearly*. Cambridge: Cambridge University Press.

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

Scovel, T. (2000). A critical review of the critical period research. *Annual Review of Applied Linguistics, 20*, 213-23.

Smith, J. (1994). Enhancing curricula for ITA development. In C.G. Madden & C.L. Myers (Eds.), *Discourse and performance of international teaching assistants* (pp. 52-62). Alexandria, VA: Teachers of English to Speakers of Other Languages.

Smith, L.E. (1992). Spread of English and issues of intelligibility. In B.B. Kachru (Ed.), *The Other Tongue: English across cultures* (pp. 75-90). Urbana, IL: University of Illinois Press.

Stake, R.E. (1995). *The art of case study research.* Thousand Oaks, CA: SAGE Publications.

Strauss, A. & Corbin, J. (1998). *Basics of qualitative research* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Taylor, D.S. (1991). Who speaks English and to whom? The question of teaching English pronunciation for global communication. *System, 19*(4), 425-435.

Turner, C.E. & Upshur, J.A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1), 49-70.

Upshur, J.A. & Turner. C.E. (1999). Systematic effects in the rating of second language speaking ability: Test method and learner discourse. *Language Testing, 16*, 84-111.

Wennerstrom, A. (2001). *The music of everyday speech.* Oxford: Oxford University Press.

Wong, R. (1987). Learner variables and prepronunication considerations in teaching pronunciation. In J. Morley (Ed.), *Current perspectives on pronunciation: practices anchored in theory* (pp. 13-28). Washington, DC: Teachers of English to Speakers of Other Languages.

Yin, R.K. (2003). *Case study research: Design and methods* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Yule, G. (1990). Reviews [Teaching English pronunciation; Current perspectives on pronunciation; Practices anchored in theory]. *System, 18*(1) 107-111.

Appendixes

Appendix A

Certificate of Ethical Acceptability

Appendix B

Non-Native Speaker Consent Form

INFORMED CONSENT FORM TO PARTICIPATE IN RESEARCH

Title: *Towards defining a valid criterion for the assessment of proficiency in the pronunciation of non-native English speaking graduate students*
Principle Investigator: Talia Isaacs, McGill University
Faculty Supervisor: Dr. Carolyn E. Turner

Purpose
This study sets out to examine whether and to what extent intelligibility is an appropriate criterion in assessing the pronunciation proficiency of non-native English speaking graduate students in the academic domain.

Procedures
Ten native and ten non-native English speaking graduate students in the Faculty of Education will be invited to participate in this study on a voluntary basis.[63]
Speech samples of the non-native speakers will be elicited using the *Test of Spoken English*, and audio recordings will be randomized on a CD. The native-speakers will be asked to rate the overall intelligibility of the speech samples and to comment on factors contributing to speech clarity. A follow-up questionnaire will be administered to all participants for background information.
The speech and questionnaire data will be transcribed and analyzed by the researcher. All nominal information will be protected for confidentiality by assigning a random identification code to each respondent in the data set. Speech samples will be erased once the data analysis is completed.

Conditions of Participation

- Native speakers will be paid $20 for an estimated one hour of their time; non-native speakers will be paid $10 for an estimated thirty minutes.
- There are no risks involved in participating in this study, other than that you may, perhaps, feel uncomfortable about having your speech recorded and assessed.
- The benefits for you in participating in the study could be personal. The study may, for instance, offer insight into those factors which contribute to the perception of clarity in speech. You may also feel satisfaction in contributing to future research in the areas of pronunciation and assessment.
- Participation in this study is strictly voluntary and will not affect your grades or the evaluation of your work in any way.
  You are free to withdraw from the study at anytime without penalty or prejudice.
- Under no circumstances will any information regarding your personal identity be disclosed. In the write-up, anonymity will be maintained through the use of pseudonyms.

I have read and understand all of the above conditions. I freely consent and voluntarily agree to participate in this study.

Name (please print): ____

Signature: ______________________________ Date: ________________________

[63] This consent form was distributed to graduate non-native speakers before it was decided that the native speaking raters would be undergraduate science students. For the undergraduate consent form, see Appendix C.

Appendix C

Native Speaker Consent Form

INFORMED CONSENT FORM TO PARTICIPATE IN RESEARCH

Title: *Towards defining a valid criterion for the assessment of proficiency in the pronunciation of non-native English speaking graduate students*

Principle Investigator: Talia Isaacs, McGill University
Faculty Supervisor: Dr. Carolyn E. Turner

Purpose
This study sets out to examine whether and to what extent intelligibility is an appropriate criterion in assessing the pronunciation proficiency of non-native English speaking graduate students in the academic domain.

Procedures
Fifteen native English speaking students registered in "Topics in organic chemistry" will be invited to participate in this study on a voluntary basis.
After filling out a short questionnaire for background information, participants will be asked to listen to 4 minute pre-recorded speech samples of 12 non-native speakers of English, rate each speaker on overall intelligibility (i.e., how much of the message they can understand), and comment on factors which they feel contribute to speech clarity.

Conditions of Participation

- You are eligible to participate in this study if English is your first language (i.e., you have had English at home before the age of three)
- You will be paid $20 for an estimated one hour of your time.
- There are no risks involved in participating in this study, other than that you may, perhaps, find the prompts somewhat repetitive.
- The benefits for you in participating in the study could be personal. The study may, for instance, offer insight into those factors which contribute to the perception of clarity in speech. You may also feel satisfaction in contributing to future research in the areas of pronunciation and assessment.
- Participation in this study is strictly voluntary and will not affect your grades or the evaluation of your work in any way.
  You are free to withdraw from the study at any time without penalty or prejudice.
- Under no circumstances will any information regarding your personal identity be disclosed.

I have read and understand all of the above conditions. I freely consent and voluntarily agree to participate in this study.

Name (please print): ______________________________________

Signature: ______________________ Date: __________________

Appendix D

Non-Native Speaker Questionnaire

Participant Questionnaire

The purpose of this questionnaire is to gather information about your background and goals as they relate to language and pronunciation. Please answer as completely as you can.

Background Information

1. Name: ____________________
2. Age: ____
3. Program/ Year of study: ________________________________

4. First language (chronologically): ________________________________
   Second language: ________________________________
   Other languages: ________________________________

5. Language(s) of schooling
   Primary: ________________________________
   Secondary: ________________________________
   CEGEP: ________________________________
   Undergraduate: ________________________________
   Graduate: ________________________________

6. Please provide any standardized test scores if known (e.g., TOEFL, TOEIC).
   Scores: ________________

Please circle the answer(s) that is/ are appropriate:

7. Period of residence in Montreal:

   1 year   2 years   3 years   4 years   5 years or more

8. Period of residence in other English speaking environments:

   N/A   1 year   2 years   3 years   4 years   5 years or more

   Please specify where: ____

9. Approximately what percentage of time do you speak English (as opposed to other languages)?

At home: 0-20%	21-40%	41-60%	61-80%	81-100%	N/A
At school:0-20%	21-40%	41-60%	61-80%	81-100%	N/A
At work: 0-20%	21-40%	41-60%	61-80%	81-100%	N/A

10. Which varieties of English have you had exposure to?

Canadian   American   British   Australian

Other (please specify):____________________

11. Have you received any explicit pronunciation instruction in English as an ESL/EFL student?

Yes   No

Please explain: ____

12. Have you ever taken a phonetics/ phonology course?

In English: Yes   No
In any other language: Yes   No

13. What is your primary goal when you speak English at school?

To sound native-like
To be understood
To get your message across
Other (please specify): __________________________

14. What are your biggest pronunciation difficulties in English?

Word stress - where to place the most emphasis in a word, (e.g., BEDroom vs. bedroom)
Sentence stress - where to place the most emphasis in a sentence, (e.g., She ate the CAKE vs. She ATE the cake)
Certain Consonants (e.g., the "th" in "bath")
Certain Vowels (e.g., the "ee" sound in "beach")
Pitch
Other (please specify): ________________________________

15. Have you observed any changes/ improvements in your pronunciation patterns over the past two years?

Yes   No

Please explain: ____

16. Do you think that pronunciation affects your ability to carry out your role as a graduate student?

Yes No

Please explain: ____

Appendix E

Native Speaker Questionnaire and Rating Scheme

Section 1 – Background Information

The purpose of this questionnaire is to gather information about your background as it relates to language and pronunciation. Please answer as completely as you can.

1. Age: ____

2. Program/ Year of study: ______________________________

3. First language (chronologically): ______________________________

   Second language: ______________________________

   Other languages: ______________________________

4. Approximately what percent of time do you speak English (as opposed to other languages) in your daily life? (Please circle one answer only).

   25% 50% 75% 100%

5. Have you ever taken a phonetics/ phonology course or had any pronunciation training?

   Yes No

If the answer is yes, please describe the context:

____

____

Section 2 – Rating the Speakers[64]

In this section, you'll be asked to rate the pronunciation of a non-native speaker of English. You will fill out this section a total of 8 times for 8 different speakers ***before*** *proceeding to Section 3.*

Speaker #1

As you listen to this person's speech for the first time, see if you can understand every single word that they say.

> **Remember:** By "understand" I mean that you are able to comprehend each word immediately so you do not have to guess at words.

On the scale below, mark approximately what percent of the speaker's words you are able to understand with an "X."

0% 25% 50% 75% 100%

Now you will hear the speech samples again. As you listen, try to identify whether any of the features listed below hindered your ability to understand the speaker's words.

□ **a. speech clarity** - the speaker:
- □ overpronounces words (articulates each syllable painstakingly)
- □ mumbles/ eats words (speech is unclear or muffled)

□ **b. rate of speech** - the speaker:
- □ speaks too fast
- □ speaks too slowly

□ **c. pitch** – the speaker's pitch:
- □ changes too often from high to low
- □ doesn't change enough/ is too monotone

□ **d. sentence rhythm** - the speaker:
- □ fails to distinguish between important and unimportant words in the sentence
- □ fails to link sounds between words (e.g., doesn't connect the "z" sound to the "a" in applez‿and oranges")

[64] In the version of the questionnaire that was given to the raters, the top and bottom margins of *Section 2* were altered so that all pronunciation features up to and including the "None" box appeared on the same page to avoid excess page flipping during the playing of the speech samples. Margins were normalized for the reproduction of this thesis based on the page formatting guidelines.

□ **e. <u>word stress</u>** – the speaker:

□ often doesn't get the syll-A-ble right

□ often doesn't distinguish between strong and weak syllables

□ **f. <u>individual consonant/ vowel sounds</u>** – the speaker:

□ substitutes problematic sounds for ones that are easier to pronounce (e.g., says "sink" instead of "think" or "heat" instead of "hit")

□ adds sounds or deletes sounds (e.g., says "sundly" instead of "suddenly" or "warem" instead of
"warm")

* □NONE

Go back to the previous page and:

1.) Rank order the top 3 features that hindered your ability to understand the speaker's words by placing a number in the big box. 1 is for the feature that most hindered your ability to understand, 2 is for the second most hindering feature, 3 is for the third one.

2.) For whatever 3 features you have rank ordered, check the most prominent problem (i.e., the one that stands out to you the most) in the small box below the letter. You may only check one option.

3.) If none of features #a-f interfered with your ability to understand the speaker's words, leave those boxes blank and mark an "X" in the NONE ***box at the bottom of the page.***

4. Are there any other features that hindered your ability to "understand" this speaker? If so, please write them below:

____

____

5. How familiar are you with this speaker's accent?

□ familiar □ somewhat familiar □ unfamiliar

6. If you think you know the speaker's first language is, write it in the blank: __________________

7. How would you rate this speaker in terms of you being able to understand?

□ very difficult □ difficult □ easy □ very easy

8. Do you think that this person's pronunciation is adequate to be a Teaching Assistant (TA) in an undergraduate level course?

□ yes □ no □ not sure

9. Please write any other comments about this person's speech below:

____

____

Section 3 – Summing up your listening/ rating experience

We're almost done! Just a few more questions to finish up. Please refer to previous ratings to help jog your memory.

1. Are there any speakers that really stand out in terms of being particularly easy to understand? If so, please list them below:

Speaker # __
Speaker # __

2. Are there any speakers that really stand out in terms of being particularly difficult to understand? If so, please list them below:

Speaker # __
Speaker # __

3. Of the pronunciation features you identified in rating the speakers, which three do you think are the most critical to understanding the speakers that you heard? Please write them in order of importance in the blanks below:

(speech clarity; rate of speech; pitch, sentence rhythm, word stress, individual sounds)

1. ______________
2. ______________
3. ______________

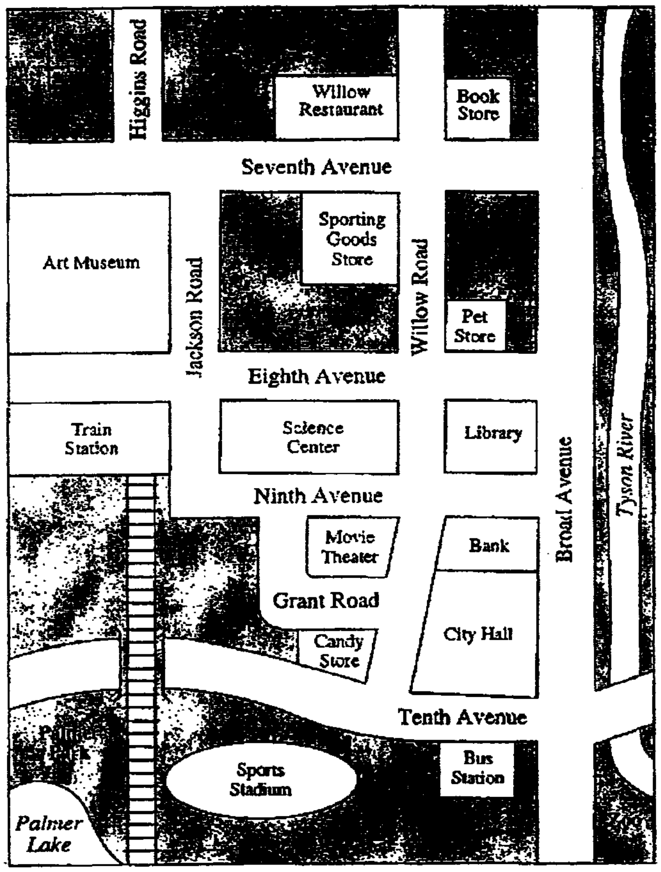
Congrats on getting through this rating session! Thank you for your help.

Appendix F

1995 Version of the Test of Spoken English[65]

[65] Reproduced by permission of Educational Testing Service, the copyright owner

Imagine that I'm a friend of yours. This is a map of a nearby town that you have suggested I visit. You will have thirty seconds to study the map. Then I'll ask you some questions about it.



1. There are several interesting places to visit in this town. Choose one place on the map you would recommend that I visit. Tell me why you recommend this place. (30 seconds)

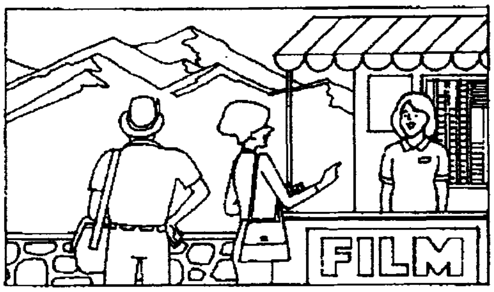
2. I am going to meet you for lunch at the Willow Restaurant. Please give me directions from the bus station to the restaurant. (30 seconds)

3. During lunch we have been discussing other restaurants. I'm interested in hearing about your favorite restaurant. Please describe it to me in as much detail as you can. (45 seconds)

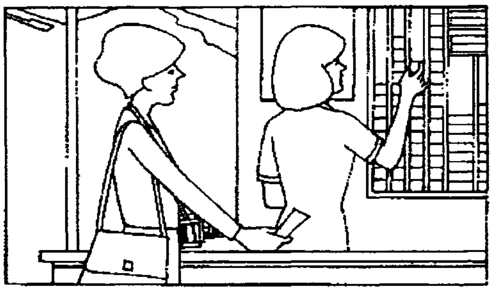
4. The city officials have proposed that the central area of this town that is between the train and bus stations be limited to public vehicles and nonmotorized traffic, such as buses and bicycles. Some people think that all areas of town should be open to private cars. Which point of view do you agree with and why? (60 seconds)

GO ON TO THE NEXT PAGE

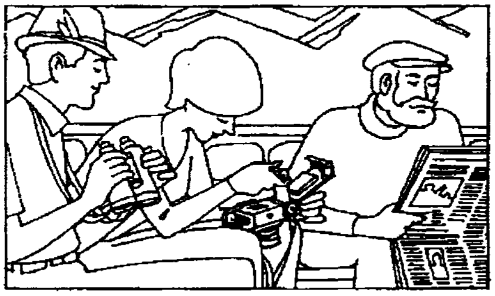
Now please look at the six pictures below. I'd like you to tell me the story that the pictures show, starting with picture number 1 and going through picture number 6. Please take one minute to look at the pictures and think about the story. Do not begin the story until I tell you to do so.



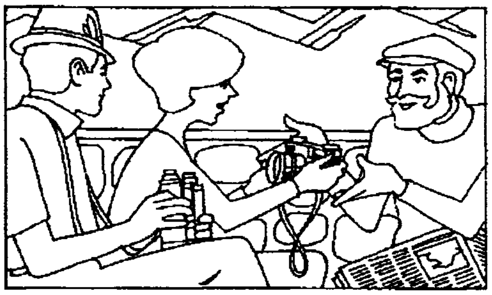
1



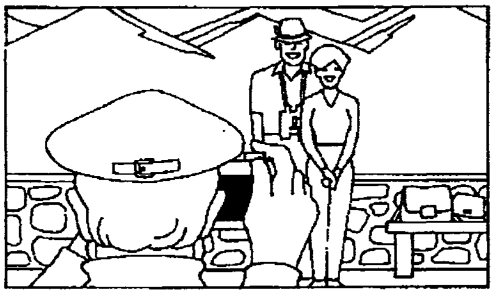
2



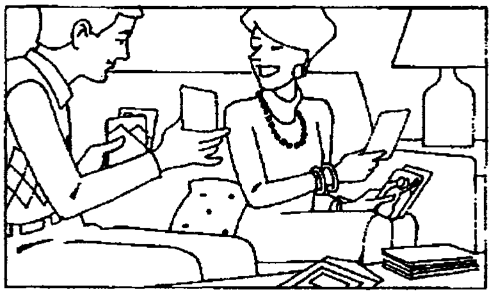
3



4



5



6

5. Now tell me the story that the pictures show. (60 seconds)

6. What would you do in this situation if the man refused to take your picture? (30 seconds)

7. Some people enjoy taking photographs when they travel to have a record of their trip. Other people prefer to make written notes about their experiences. What do you think are the advantages and disadvantages of each method? (60 seconds)

8. Imagine that you have been on vacation and have taken many photographs. You take your film to the store to be developed. When you pick up your order from the store and return home, you discover that you've been given someone else's photographs. Call the store and complain to the manager about the problem. (60 seconds)

GO ON TO THE NEXT PAGE

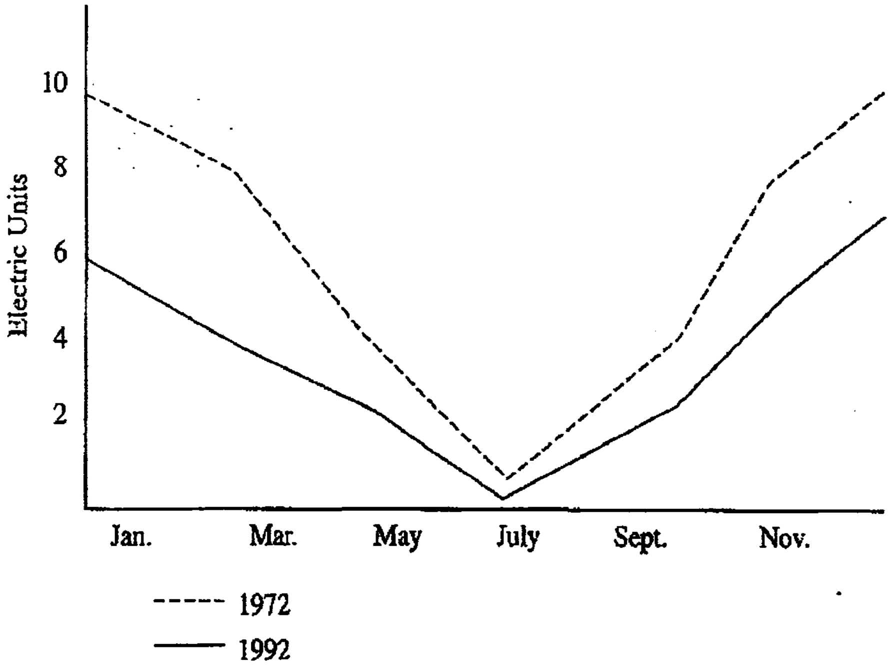
Now I'd like to hear your ideas about several topics. Be sure to say as much as you can in responding to each question. After I ask each question, you may take a few seconds to prepare your answer, and then begin speaking when you're ready.

9. I know very little about your field of study but am interested in learning more about it. Tell me about a typical research project that someone in your field might carry out. (60 seconds)

---

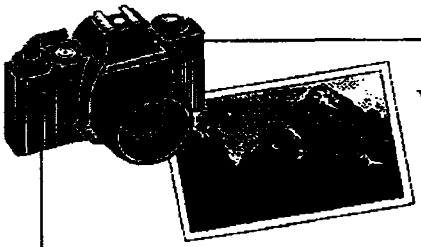
10. The graph below represents the number of electric units used to heat a home in North America in 1972 and in 1992. Describe the information given in the graph. (45 seconds)

ELECTRIC UNITS USED FOR HEATING A HOME



11. Tell me what you think are the possible causes for the differences in electricity used in the two years. (45 seconds)

12. Imagine that you are the president of a local photography club and are at the monthly meeting. At the last meeting you distributed some information about a photography contest that will be held in the area. Since then there have been some changes to the schedule. Remind the club members of the important details and point out to them the changes in the notice. In your presentation do not just read the information printed, but present it as if you were talking to a group of people. You will now have one minute to plan your presentation. Do not start speaking until I tell you to do so.



WOODLANDS NATURE CENTER PHOTOGRAPHY CONTEST

Categories:*	A. Wildlife B. Landscape
Important Dates:	Deadline for entries: Friday, May 10, ~~5:00~~ 3:00 p.m. Nature Center Main Office Judging: Monday, May 13 Exhibition: May 15–June ~~15~~ 30 Nelson Gallery
Judges:	Susan Milton ~~Judith Morgan, Associate~~ Professor, Wilson Art School Mark Stewart, Fine Arts Director, Metropolitan Museum
Fee:	1 entry $5.00 2 or more entries $8.00
Reception:	Friday, May 17, 5:00–7:00 p.m., Nelson Gallery Room B Public invited

*Limit 2 photos in each category

(90 seconds)

STOP

Appendix G

Transcription Symbols and Color-Coding

Transcription symbols - suprasegmentals

Textual symbols for stress, rhythm, pitch, and intonation
' primary word stress
underlined multi-syllable word where there is no clear main stress
• words emphasized in the sentence (sentence rhythm)
◇ lack of distinction between important and unimportant words
subscripted low pitch
superscripted high pitch
subscripted indented H→L tone sliding
superscripted indented L→H high rise intonation
-= =- monotone, no inflection
↑+3 ↓-6 pitch rises ↑ or falls ↓ by the musical pitch internal indicated

Textual symbols for pauses and tempo
»hurry up«
«slow down»
_ brief unmeasured pause (shorter than one second)
(1.2.3) pause measured in seconds when it's one full second or more
- consonant/ vowel sound prolonged at same pitch
! staccato
/ last two sounds detached (no linking)[66]
‿ legato and/ or linking from one sound to another

Other textual symbols
((cough))

Transcription symbols - segmentals

Textual symbols for individual sounds besides IPA symbols
# sound deletion

Intelligibility coding

Color-coding
green – unusual pronunciation, but does not affect intelligibility
red – unintelligible pronunciation, or results in unintelligibility

[66] ! and / were often used in tandem and are, for the most part, interchangeable. Subtle differences do exist in the use of the symbols, however. ! tended to be used if word or sound was aborted early and abruptly, whereas / was used to mark the more discrete lack of linking between sounds. !/ indicates an especially strong effect.

Appendix H

Non-Native Speaker Descriptive Data

Non-Native Speaker Descriptive Data

Speaker	Sex	Program of Study	L1(s)	No. of Langs. Spoken	1.Pronunciatn. instruct.ion 2. Phonlogycours	Period of Residence in Montreal	Res. in Other EngSpeaking Place	Teaching Experience	TA Experience	Plans after graduation
A	Male	2nd Language	Korean	3	Both	1 year	2 years	EFL	-	Academia/ Teach ESL
B	Female	2nd Language	Mandarin	3	Neither	3 years	-	ESL/ EFL	-	Academia/ Teach ESL
*C	Female	Ed Psych	Korean	2	#1 Only	2 years	-	-	-	Teach EFL
**D	Female	2nd Language	Mandarin/ Taiwanese	4	#2 Only	2 years	-	EFL	-	ESL/ EFL
*E	Female	2nd Language	Mandarin	2	Both	2 years	-	EFL	-	Teach EFL
*F	Male	2nd Language	Japanese	5	#2 Only	1 year	1 year	EFL	-	Teach EFL
*G	Male	Kinesiology	Korean	2	Both	2 years	-	Physiology	Yes	Fitness
H	Female	2nd Language	Japanese	2	#1 Only	2 years	-	EFL	-	Teach EFL
I	Female	2nd Language	Mandarin	2	#1 Only	2 years	-	EFL	-	Teach EFL
J	Male	2nd Language	Japanese	3	Both	2 years	1 year	EFL	-	Teach EFL
*K	Male	Ed Psych	Serbo-Croatian	2	#1 Only	1 year	5 years +	Special Ed	-	Academia
L	Female	Curriculum	Malay	4	#1 Only	1 year	-	Science	-	Teacher Training
*M	Female	Curriculum	Indonesian/ Javanese	3	#1 Only	1 year	-	Science	-	Teach Arabic
*N	Male	2nd Language	Spanish	3	Both	5 years +	-	ESL/ EFL	Yes	Academia/ Teacher Training
O	Female	2nd Language	Japanese	2	#2 Only	2 years	1 year	EFL	-	EFL
P	Female	Curriculum	Sundanese	3	#1 Only	1 year	-	EFL	-	Teacher Training
Q	Female	2nd Language	Mandarin	4	#1 Only	2 years	-	ESL/ EFL	-	EFL
*R	Female	Culture&Values	French	4	#1 Only	5 years +	-	French FL	-	Academia
S	Female	Culture&Values	Japanese	3	Neither	2 years	3 years	EFL	Yes	NGO

*Indicates that the speaker was included in the rating session. **Indicates that this was the "Practice Speaker" in the rating session.