Issues and Arguments in the Measurement of Second Language Pronunciation

Talia Isaacs, Department of Integrated Studies in Education

McGill University, Montreal

October, 2010

A thesis submitted to McGill University in partial fulfillment

of the requirements of the degree of Doctor of Philosophy

Acknowledgments

During my Master's degree at McGill University, I forged valuable ties in Montreal, and was lucky to build on this tremendous network of support during my Ph.D. Dr. Carolyn Turner, my advisor for both my Master's and my Ph.D., has fostered the development of my voice as a researcher by giving me the scope I needed to think out of the box while, at the same time, keeping me on track. Carolyn has modeled mentorship, and I will always bear the stamp of her student through my interest in mixed methods research and rating scales. I am delighted to have worked with Dr. Pavel Trofimovich and Dr. Ron Thomson on the manuscripts included in this thesis. Pavel has catalyzed the maturing of my scholarly writing and has helped me integrate my musical background into my research in new and exciting ways. Presenting with Ron at annual conferences is always a highlight of my year. I repeatedly marvel at how we mostly see eye-to-eye and complement each other so well given our different strengths and perspectives.

I have also had the privilege of forging strong relationships with my committee members, Dr. Mela Sarkar, Dr. Lise Winer, and Dr. Tracey Derwing. Crucial moments in my doctoral journey are punctuated by vivid memories of our interactions. These include reflecting on the drudgery of comprehensive exam writing while inhaling therapeutic chai vapors at Mela's Annual Tea Party; coming into a meeting with Lise on a stark winter day drained and directionless but leaving energized and with my doctoral proposal clear in my mind; and finally, working alongside Tracey on the plane after AAAL in Costa Mesa and, more recently, receiving timely job interview and post-Ph.D. strategies in Edmonton and Montreal. These relationships have significantly enhanced the quality of my Ph.D. experience.

Thanks are due to my Ph.D. colleagues, Beverly Baker, Christian Colby-Kelly, Candace Farris, Pamela Gunning, Heike Neumann, Dr. May Tan, and Dr. Jing Wang, all of whom I can count on for high-quality feedback. I am also indebted to Concordia University's Research Group for including me as part of the research community and providing me with an incentive to maintain human contact during stressful periods. I have also benefited greatly though working with Dr. Carolyn Turner, Dr. Michel Laurier, and Dr. Norman Segalowitz on the H-CALM assessment team, and have a feeling that the research mantra that "each grant application is a springboard for the next," which I have now adopted as my own, will stand me in good stead post-Ph.D. (thank you, Norman). Finally, Dr. Gad Lim, a new member of my inner circle, has laughed at my tacky applied linguistics jokes, mused over cryptic expressions of frustration, and been steady, sturdy, and always candid, especially during "horseraces."

I am most grateful for the financial support that I received during my doctorate, including a Canada Graduate Scholarship from the Social Sciences and Humanities Research Council of Canada, the Sir James Lougheed Award of Distinction from Alberta Scholarship Programs, and the J. W. McConnell Memorial Fellowship and Recruitment Excellence Fellowship from McGill University.

Throughout my Ph.D., I have leaned heavily on my family and friends. Honey, Eddy, and Nili, thank you for believing in me—I could not have gotten here without you. And Honey, it goes without saying that your "objective eye" is invaluable and gives me a fresh perspective. Thank you for reading every word of my thesis. Rachelle and Kathleen, I have grown with you both through grant writing and other endeavors and am so grateful for your friendship. Finally, Martin, by challenging my assumptions, you keep me honest and always thinking and have been there for me when it has counted the most.

Contributions of Authors

The McGill University "thesis preparation" guidelines provide candidates (i.e., examinees) with the option of preparing a manuscript-based thesis, comprised of a collection of articles that are part of the same overall program of research, in lieu of a "traditional" thesis (www.mcgill.ca/gps/students/thesis/programs/guidelines/preparation). As specified in the McGill guidelines, a general introduction and conclusion to the thesis and transition texts between articles are necessary to support the cohesive nature of the thesis and the logical progression between the articles. ***The manuscript-based thesis option has been pursued in this thesis.*** The three included manuscripts are currently in press (Study 1), under revision (Study 2), or under review (Study 3) in peer-reviewed journals.

The McGill University guidelines also specify that the candidate may include co-authored manuscripts as part of the thesis, provided that he/she is the primary author (i.e., has made the most substantial contribution to each paper). In the case of co-authored manuscripts, the onus is on the candidate to clearly describe the responsibilities of all authors involved. ***All three manuscripts included in this thesis are co-authored. I have contributed to each manuscript as primary author***. Studies 1 and 3 were co-authored with Dr. Pavel Trofimovich of Concordia University, initially a member of my doctoral advisory committee with whom I subsequently collaborated. Study 2 was co-authored with Dr. Ron Thomson of Brock University, with whom I initially began collaborating while he was completing his doctorate at the University of Alberta.

I actively contributed to conceptualizing all three papers in this thesis. The initial idea for the papers stemmed from issues I felt needed to be addressed with some urgency. In Study 1, I sought to confirm my intuitions about musicians' sensitivity to second

language (L2) speech, and Dr. Trofimovich guided my orientation in research on phonological memory and attention control in a reading course at Concordia University (not required for degree completion) that eventually became the basis for this study. In Study 2, I was motivated to explore whether other raters would find 9-point scales as unwieldy and difficult to manage as I had when evaluating L2 pronunciation. Finally, having reflected on the limitations of numerical rating scales, in Study 3, I strove to describe with more precision the linguistic factors that most clearly distinguish between different levels of L2 comprehensibility in an empirically-derived scale. These papers are all a logical extension of my Master's thesis, which centered on the need to more precisely define the constructs we aim to measure in the context of L2 pronunciation research (Isaacs, 2008, in press). In addition to serving as the "assessment expert" for all three papers, I drew on my mixed-methods background in Studies 2 and 3 by drawing on different but complementary data (i.e., quantitative and qualitative) in order to better address the research problem (see Creswell & Plano-Clark, 2007).

The speech data in Studies 1 and 3 were collected as part of an earlier study by Dr. Trofimovich (Trofimovich, Gatbonton, & Segalowitz, 2007) that I obtained permission to use, and Dr. Thomson collected and edited the speech data for Study 2. I recruited the raters for all three studies, including determining the selection criteria and matching participants. I prepared data elicitation materials for raters, including written questionnaires (Studies 1–3), think-aloud protocol instructions and oral scripts (Studies 2 and 3), and post-task interview guidelines (Study 2). Once these instruments were developed, my co-authors helped refine them. I trained research assistants on data transcription procedures for Study 2, on how to elicit the introspective data from the teachers in Study 3, and on how to apply the coding scheme I had developed for

intracoder reliability, also in Study 3. Dr. Trofimovich and I, who jointly brainstormed and finalized the speech measures included in Study 3, participated in the initial analysis and second coding of a subset of measures. The remaining measures were analyzed by a team of research assistants that Dr. Trofimovich trained.

Following the research assistants' transcriptions of the L2 speech samples (Studies 1–3) and rater comments (Study 2), I verified the transcripts and conducted data entry and analyses, including conventional statistics (i.e., from classical test theory) for all studies, Rasch category probability plots for Study 2, and qualitative coding and content analysis for Studies 2 and 3. Finally, I wrote the initial draft of all manuscripts as the primary author, including the L2 comprehensibility scale in Study 3, and integrated my co-authors' feedback during the editing process. I have served as the corresponding author for all submitted manuscripts, implemented the reviewers' changes for Study 1, and have kept my co-authors updated on all editorial correspondence.

In summary, I have taken a lead role in conceptualizing the studies, preparing the instruments, conducting the analyses, and independently writing the initial draft of the manuscripts. Although my co-authors provided the speech data, I was responsible for either collecting the rater data (Study 2) or training research assistants to do so (Studies 1 and 3), and was involved in rater recruitment in all three studies. In addition to the co-authored manuscripts in this thesis, I have a strong track record of independent research. I have produced four single-authored, peer-reviewed journal articles or book chapters and one book review (four in print, one in press). These publication experiences have enabled me to collaborate effectively.

Abstract

This thesis examines systematic sources of variance in raters' judgments of second (L2) language speech, including rater cognitive and experience variables, rating scale properties, and characteristics of the speech, in order to better understand influences on raters' scoring decisions. The thesis culminates in the development of an empirically-based L2 comprehensibility scale that describes, with greater precision, the quality of speech that is characteristic at different comprehensibility levels.

Study 1 examines the effect of individual differences in raters' cognitive abilities on their ratings of L2 speech. Thirty music majors and 30 non-music majors rated 40 L2 speech samples for comprehensibility, accentedness, and fluency and were additionally assessed for musical ability, phonological memory, and attention control. Results showed that music majors assigned significantly lower ratings than non-music majors solely for accentedness, particularly for low ability learners. However, phonological memory and attention control did not influence their ratings.

Study 2 examines the effects of two additional sources of variance—rating scale length and rater experience—on raters' judgments of L2 comprehensibility, accentedness, and fluency. Twenty experienced and 20 novice raters judged 38 L2 speech samples using 5-point or 9-point numerical rating scales. In addition, raters' perceptions of the rating process were elicited through verbal protocols and interviews. Results showed that experienced and novice raters achieved high consensus about the highest and lowest scoring L2 speakers but had difficulty differentiating between scale levels in the absence of guidance from the rating instrument.

Finally, the goal of ***Study 3*** was to construct an L2 comprehensibility scale rooted in raters' perspectives of influences on their judgments, and characteristics of the L2

speech. To this end, 19 speech measures used to analyze 40 L2 speech samples were examined in relation to 60 raters' mean L2 comprehensibility ratings and three ESL teachers' indications of their most salient scoring criteria. Overall, a wide range of measures contributed to listeners' comprehensibility judgments, with vocabulary and fluency measures distinguishing between low-level learners, grammatical and discourse-level measures distinguishing between high-level learners, and word stress distinguishing between all levels. Taken together, these papers advance our understanding of raters' perspectives in L2 pronunciation assessment.

Résumé

Cette thèse examine les sources de variation systématique dans le jugement des évaluateurs de la production orale en langue seconde (L2)—incluant les variables cognitives et l'expérience des évaluateurs, les propriétés des échelles de compétences, et les caractéristiques du discours—afin de mieux comprendre les influences sur le jugement des évaluateurs. La thèse se termine par l'élaboration d'une échelle empirique de compréhensibilité en L2 qui décrit, avec plus de précision, la qualité linguistique qui caractérise des niveaux différents de compréhensibilité.

La 1ère étude évalue l'impact des différences individuelles dans les capacités cognitives des évaluateurs sur leurs jugements de productions orales en L2. Trente étudiants faisant leur baccalauréat en musique et 30 étudiants inscrits à d'autres facultés ont évalué 40 échantillons d'un discours L2 en fonction de sa compréhensibilité, la perception de son accent, et sa fluidité. De plus, les participants ont été évalués pour leur aptitude musicale, leur mémoire phonologique, et leur contrôle de l'attention. Les résultats démontrent que les étudiants en musique évaluent le volet de la perception de l'accent des échantillons du discours de façon plus sévère, surtout pour les apprenants de L2 de faible niveau. Toutefois, la mémoire phonologique et le contrôle de l'attention n'exercent aucune influence sur l'évaluation des évaluateurs.

La 2e étude examine les effets de deux sources de variation additionnelles—la longueur de l'échelle de compétence et l'expérience des évaluateurs—sur l'estimation des évaluateurs de la compréhensibilité, de la perception de l'accent, et de la fluidité du discours L2. Vingt évaluateurs expérimentés et 20 évaluateurs novices ont jugé 38 échantillons d'un discours L2 en utilisant des échelles numériques à 5-points ou à 9-points. De plus, le cheminement cognitif qui à mené les évaluateurs à leur choix de score

ont été obtenues par des protocoles verbaux et des entretiens. Les résultats montrent que les évaluateurs expérimentés et novices sont parvenus à un consensus sur les échantillons de discours de plus haut niveau et de plus bas niveau, mais ont éprouvé des difficultés à différencier entre les niveaux de l'échelle en l'absence de directives de l'instrument.

Enfin, l'objectif de ***la 3e étude*** vise à construire une échelle de compréhensibilité en L2 ancrée dans les commentaires des évaluateurs sur les facteurs qui ont influencés leurs jugements et, en fonction des caractéristiques du discours L2. À ces fins, 40 échantillons du discours L2, analysés au moyen de 19 mesures linguistiques, ont été examinés par rapport aux jugements de compréhensibilité de 60 évaluateurs, et par rapport aux critères d'évaluation de trois enseignants d'anglais langue seconde. Les résultats montrent qu'un grand éventail de mesures contribue aux jugements de compréhensibilité des évaluateurs. Les apprenants de bas niveau se différencient par le biais des mesures de vocabulaire et de fluidité tandis que les apprenants de haut niveau se différencient plutôt par les mesures de grammaire et de discours et enfin, l'accent de mot permet de différencier les apprenants de tous niveaux confondus. L'ensemble de ces résultats avancent notre compréhension des perspectives des évaluateurs dans l'évaluation de la prononciation.

Table of Contents

List of Tables and Figures

Study 1

Study 2

Study 3

Definitions of Key Terms Used in this Thesis

For the purpose of this thesis, the key terms are defined as follows:

Comprehensibility – Listeners' perceptions of how easily they understand a second language (L2) utterance. This measure is quantified by eliciting listeners' scalar ratings of ease/difficulty of understanding a given speech sample (see Munro & Derwing, 1999).

Intelligibility – The extent to which listeners are able to understand L2 speech (Munro & Derwing, 1999). This construct is most often operationalized by the proportion of an L2 speaker's utterance that listeners are accurately able to transcribe, although other measures of intelligibility have also been used (e.g., listener response accuracy to true/false questions; see Derwing & Munro, 2009a).

Accentedness – Listeners' perceptions of how closely the pronunciation of an L2 utterance resembles that of a native speaker of North American English. This measure is quantified by eliciting listeners' scalar ratings of the degree of accent of a given speech sample (see Munro & Derwing, 1999).

Fluency – Listeners' perceptions of the how smoothly and rapidly an L2 utterance is delivered. This measure is quantified by eliciting listeners' scalar ratings of fluency. The term "fluency measures" refers to temporal measures used to analyze L2 speech samples using speech editing software (e.g., pause length, mean length of run, etc.; see Derwing, Rossiter, Munro, & Thomson, 2004).

Assessment – The process of gathering information about a test-taker's ability on the variable of interest. "Tests" and "evaluations" are specific types of assessments (see Bachman and Palmer, 2010).

Chapter 1 – General Introduction

In 1957, the English linguist, J. R. Firth, famously wrote, "you shall know a word by the company it keeps" (p. 11). A quick perusal of the past several decades of second language (L2) pronunciation research reveals that the term "pronunciation" has kept close company with "neglect" (e.g., Derwing & Munro, 2009b; Gilbert, 1994; Isaacs, 2009; Jenkins, 1998; Neri, Cucchiarini, Strik, & Boves, 2002; Lord, 2008; Rogerson & Gilbert, 1990). This disparaging association generally refers to the devaluation of pronunciation by some communicative proponents—a practice which started in the late 1960s—and its resulting de-emphasis in the ESL classroom, the effects of which are still being felt today. To counter this view, Morley (1991) has argued that intelligible pronunciation is an indispensible part of communicative competence and needs to be reintegrated into the "instructional equation" (p. 488). This perspective has been embraced by both L2 pronunciation researchers, and teaching professionals who are active in the IATEFL "Pronunciation Special Interest Group" or the TESOL "Speech, pronunciation, and listening interest section." However, repercussions of the neglect of pronunciation are still being felt in teacher training and pedagogical practice (see Breitkreutz, Derwing, & Rossiter, 2001; Gilbert, 2010; MacDonald, 2002).

Although the subject of L2 pronunciation *teaching* conjures up reference to neglect, there is at least a body of literature documenting this neglect. Not the same can be said about L2 pronunciation *assessment*. Rare chapters on the topic by Goodwin, Brinton, and Celce-Murcia (1994) and Celce-Murcia, Brinton, and Goodwin (1996) stand alone in underscoring that assessment issues have been paid little attention within the L2 pronunciation teaching literature. In addition, only a handful of research articles have specifically interfaced L2 pronunciation with assessment (e.g., Harding, 2008, in press;

Isaacs, 2008, in press; Jenkins, 2006; Kang, 2008; Koren, 1995; Levis, 2006; Munro, 2008; Munro and Derwing, 1994; Szpyra-Kozłowska, Frankiewicz, Nowacka, Stadnicka, 2005).

The reality is that L2 pronunciation assessment has virtually been dropped from the research agenda since the publication of Lado's *Language Testing* (1961) nearly five decades ago. Although there is no mass reversal of this trend, a glimmer of hope is apparent in the recent publication of a paper on automated pronunciation scoring in the prominent assessment journal, *Language Testing* (Franco et al., 2010), which is only the second pronunciation-focused article that has been published since the inception of the journal in 1984. More promisingly, pronunciation was well represented at the 2009 *Language Testing Research Colloquium* in Denver, CO, with three papers presented on the topic including one I co-presented with Ron Thomson (Isaacs & Thomson). Spearheaded by the work in my Master's, ***the primary goal of my research*** has been to reinvigorate the conversation on L2 pronunciation assessment by making key issues relevant to diverse academic audiences. The three co-authored papers that make up this thesis break new ground and are part of this objective.

In light of SLA and psycholinguistic research on L2 learners' *cognitive abilities* and language assessment research on *rater* background characteristics, ***Study 1*** with co-author Pavel Trofimovich examines the influence of *raters' cognitive abilities* on their assessments of L2 speech. The central issue examined is whether individual differences in raters' musical ability, phonological memory, and attention control—factors extraneous to the L2 speaking ability that is being measured—are possible sources of rater bias that could threaten the validity of raters' subjective judgments of speech.

Study 2 with co-author Ron Thomson examines the increasingly pervasive use of 9-point numerical rating scales in L2 pronunciation research in light of arguments from accent scaling research that a scale with at least nine levels is necessary to prevent a ceiling effect (Southwood & Flege, 1999). However, there are contrary indications from the L2 assessment literature that raters have difficulty distinguishing between nine levels of a rating scale (Alderson, 1991). The perspectives and scoring behavior of experienced (ESL teacher) raters versus novice (non-ESL teacher) raters are examined using a mixed-methods approach.

Finally, ***Study 3*** with co-author Pavel Trofimovich centers on comprehensibility—a major construct in L2 pronunciation research that is compatible with the instructional goal of helping L2 learners become more easily understandable to their interlocutors. Although comprehensibility has been modeled in several L2 oral proficiency scales, shortcomings of existing scales (e.g., the use of vague or relativistic descriptors) reflect a poor theoretical basis for understanding the role of comprehensibility within the broader construct of L2 oral proficiency, and a poor empirical basis for understanding the way comprehensibility manifests at different L2 ability levels. The major contribution of this study is, therefore, to derive a data-driven L2 comprehensibility scale. Scale development is informed by both raters' perspectives of the most salient influences on their judgments, and the linguistic measures used to analyze the L2 speech samples that most efficiently distinguish between low, intermediate, and L2 comprehensibility ability levels.

Taken together, the studies in this thesis examine systematic sources of variance in ratings of L2 comprehensibility, accentedness, and fluency, in order to reveal construct-relevant and construct-irrelevant influences on raters' scalar judgments. In the

next section, this research will be situated in a broader discussion on validity by drawing on the views of language testers and validity theorists.

Overarching Review of the Literature

"Construct" and "Validity:" Keywords that Need to be Examined in Tandem

Messick's (1989) unitary validity framework, which includes value implications and the social consequences of testing, has been highly influential in current thinking on educational measurement and language assessment, as is attested by annual awards given in his honor in both fields. Several decades after the unitary validity framework was first proposed, the issue of the extent to which different stakeholders (e.g., test developers) should bear responsibility for the social consequences arising from test use and misuse is still the subject of lively discussion and debate (Fulcher, 2009; Messick, 1998b), as is the notion of whether the concept of "fairness" should be subsumed under an expanded notion of validity (Kane, 2010; Shepard, 1997). In fact, Messick's integration of "all kinds of validity…under one giant umbrella," a major reason that his validity framework has been so influential (Markus, 1998), has also been a source of criticism (Borsboom, Mellenbergh, & van Heerden, 2004, p. 1063). The main concern is that the framework is impractical for test development and validation and leaves practitioners with little concrete guidance on procedures for building validity arguments (i.e., to support the intended uses of the test), and for amassing validity evidence (Brennen, 1998; Lissitz & Samuelsen, 2007; Xi, 2008).

Despite the presence of an active field of validity theorists, Messick's (1989) model remains the dominant view of validity in the language assessment literature. This status is largely due to exposure brought about by the work of Bachman (1990) and Bachman and Palmer (1996), whom McNamara (2006) refers to as "the bearer of the

Messick legacy, a real Aeneas of our field" (p. 35) and the "true heirs of Messick" (p. 32), respectively, and McNamara will likely need to come up with a new heroic reference for Bachman and Palmer's newest book (2010). In particular, Bachman and Palmer's (1996) test usefulness framework, which has rendered Messick's abstract work more accessible to language testers (Xi, 2008), has influenced L2 test development and the external evaluation of tests (e.g., Chun, 2006; Fox & Fraser, 2009; Saville, 2003; Weigle, 2002). Thus, any current discussion on construct validity in language assessment, regardless of whether it is mediated by the work of Bachman (1990) or Bachman and Palmer (1996, 2010), is likely to be rooted in the views of Messick (e.g., Kunnan, 2008; McNamara & Roever, 2006). The three papers that make up this thesis, which center on issues of construct validity in the context of L2 pronunciation research, are no exception.

The Goal of the Thesis and Situating the Papers in a Broader Discussion on Validity

The overarching goal of this thesis is to gain a deeper understanding of major perceptual constructs in L2 pronunciation research (i.e., constructs that are defined on the basis of the way they are perceived by a listener or group of listeners). This is broadly accomplished by examining systematic effects of rater background characteristics, rating scale properties, and L2 discourse characteristics on raters' scalar judgments of L2 comprehensibility, accentedness, and fluency. This overarching literature review draws on the perspectives of validity theorists to inform the discussion on systematic sources of variance in the rating process. It then applies these perspectives to the current state of affairs in L2 pronunciation research as part of the backdrop for understanding the papers that make up this thesis. Following this review, each paper is presented in turn and focuses on different aspects of the rating process.

Study 1 investigates the effects of individual differences in raters' musical ability, phonological memory, and attention control on raters' judgments of L2 speech using quantitative methods. To my knowledge, these cognitive variables have only been examined in relation to indicators of L2 learners' target language development or performance (e.g., Slevc & Miyake, 2006; Trofimovich, Ammar, & Gatbonton, 2007a), but not with respect to raters and possible influences on their scoring decisions.

Study 2 examines rating scale use as a function of another rater background characteristic: ESL teaching experience. An innovation of this study is the combination of analytical techniques it brings together to better understand the complex phenomena of interest, the listening and rating process. Rater processes and reasons for scoring as they do cannot be directly observed, and their verbal accounts while doing the activity are inevitably incomplete (Eriksson & Simon, 1993), so there is a need to draw on multiple sources of evidence so the strength of one method can compensate for the weaknesses of the other (Teddlie & Tashakkori, 2009). In particular, statistics based on classical test theory are triangulated with qualitative evidence (rater comments from verbal protocol and interview data) and Rasch analysis (Rasch category probability plots), for the purpose of examining raters' scoring behavior, gauging their perceptions of the rating process, and ultimately, revisiting conventions to do with rating procedures in L2 pronunciation research.

Finally, ***Study 3*** adopts a mixed-methods approach to probe the linguistic dimensions that underlie raters' L2 comprehensibility judgments. The study culminates in the development of an empirically-based L2 comprehensibility scale, designed to reflect both the quality of speech manifested in L2 learners' speaking performances (Fulcher,

2008), and the linguistic dimensions that appear to have the most bearing on raters' scoring decisions (Brown, Iwashita, & McNamara, 2005).

Although comprehensibility has emerged as a major concept in L2 pronunciation research, Study 3 presents the first empirical L2 comprehensibility scale, to my knowledge, that has been developed. The attempt is to "deconstruct" comprehensibility (which has often been used as an umbrella term) in elaborated rating descriptors using the criteria that are most salient for making level distinctions. Before presenting the three papers in this thesis in sequence, however, it is useful to first consider insights from the educational measurement and language assessment literature on how to treat various sources on variability in the rating process (i.e., as a source of interest or a measurement nuisance) and examine implications for construct definition and construct validity.

Variability in the Rating Process: A Proverbial "Fact of Life" (McNamara, 1996, p. 127)

Variability is an integral part of the rating process (Galaczi & ffrench, 2007), since ratings involve both test-takers and raters who vary on myriad background characteristics (i.e., are not homogeneous), including cognitive and experience variables, and who may additionally interact with the task or rating scale in different ways to produce a performance or score (Upshur & Turner, 1999). A manifestation of variability in L2 learner performance on the trait being measured is desirable, so that different levels of L2 learners' ability can be differentiated and reflected in score descriptors and level distinctions. However, extraneous variables that are not desirable for measurement may also be reflected in the score, and this may pose problems for score interpretation from a psychometric point of view.

There is some dispute in the L2 assessment literature as to whether some sources of variance in the rating process, such as the task, constitute measurement error

(Bachman, Lynch, & Mason, 1995), or, rather, lie "at the heart of the study of the L2 construct" (Deville & Chalhoub-Deville, 2006, p. 16). The former view, which follows from a psychometric perspective, holds that the L2 ability being measured can be disentangled from the context of the assessment. A concrete manifestation of this is the multitrait-multimethod approach, for example, which separates the trait from the method used to obtain it (e.g., Bachman & Palmer, 1982). On the other hand, the latter view, which stems from a social interaction perspective, holds that the L2 ability and assessment context are inseparable, since assessments are locally-situated (Chalhoub-Deville, 2003). A concrete manifestation of this is task-based rating scales, such as Upshur and Turner's "empirically-derived, binary-choice, boundary-definition rating (EBB) scales" (1995, p. 6), which reflect salient aspects of a particular task in the descriptors and may even make direct reference to the task.

The "persistent problem" (Bachman, 2007, p. 42) of how to relate the L2 ability to the context when "language is both the instrument and the object of measurement" (Bachman, 1990, p. 2) is not resolved in this thesis. In fact, this remains one of the "fundamental considerations in language testing" three decades after the publication of Bachman's (1990) theoretical volume by that title (see also Skehan, 1998). A major source of debate is whether to dismiss variability arising from test-taker by task interactions as noise in the data (e.g., Messick, 1989), or rather to embrace it as illuminating different dimensions of the construct (i.e., those that are stable versus those that fluctuate across contexts; see Chalhoub-Deville, 1996; Fulcher & Davidson, 2007; Moss, 1996, 2003). Such considerations have implications for the extent to which the task is embedded in the definition of the construct, including in test specifications and rating scales, and on the generalizability of the scores beyond the research or assessment context

where the performance samples were elicited (e.g., whether results can be extrapolated to performance on other tasks and task-types in the target language use domain; see Bachman & Palmer, 1996; Fulcher, Davidson, & Kemp, 2010).

Before exploring these issues further, it is important to first move beyond the research setting of these papers and consider what is at stake in real-world assessment contexts, when ratings of L2 speech form the basis of decision-making that has consequences for stakeholders. To this end, Messick's (1989) discussion of two types of threats to construct validity will be examined. These threats, which are operative in all assessment situations (Messick, 1996), interfere with the projection of the test-taker's ability on the construct being measured that is reflected in the test score, and could undermine the trustworthiness of the inferences that are made on the basis of these scores. Thus, threats to construct validity, which need to be minimized through rigorous, ongoing inquiry, are a focal point of test validation efforts (Hubley & Zumbo, 1996) and germane to any discussion of constructs and variability in the rating process.

In addition, Messick's views have influenced both sides of the debate on the separability of ability from context, since even those unaligned with Messick's perspective acknowledge carrying"'positivist' baggage" as researchers (Deville & Chalhoub-Deville, 2006, p. 10). In this respect, an examination of Messick's views and the psychometric tradition from which they stem is foundational to understanding different perspectives within language testing.

Construct Underrepresentation

The first threat to construct validity is construct underrepresentation, or the notion that important dimensions of the construct are not being captured in the assessment, which is too narrowly focused. For example, an L2 pronunciation speaking test could

assess test-takers' productions of segmental (i.e., vowel and consonant sounds) and suprasegmental (e.g., word stress, rhythm, intonation) aspects of speech but fail to include other dimensions relevant to the construct of interest (however defined) that, if present, would have enabled test-takers to demonstrate their ability in that area (e.g., voice quality).

Construct underrepresentation inevitably arises when constructs are operationalized in rating scales, since scale descriptors necessarily oversimplify the complex processes involved in L2 acquisition or task performance that they aim to represent (Brindley, 1998; Lumley, 2005). In the case of multidimensional constructs, scale descriptors may fall short of reflecting the myriad factors (and interactions between these factors) that raters attend to when arriving at scoring decisions, and some criteria that raters heed may not even be reflected in the rating instrument (Douglas, 1994). To add to this complexity, in the case of data-driven rating scales, different dimensions of the construct are likely to be emphasized as a function of the L2 performance samples and sampling of tasks used to generate the scale, in addition to the group of scale developers who decide which dimensions to emphasize in the descriptors (Chalhoub-Deville, 1996; Turner & Upshur, 2002; Upshur & Turner, 1999). Finally, the practical consideration of needing to provide raters with a user-friendly instrument with a manageable number of assessment criteria appears to be at odds with representing the construct comprehensively in descriptors.

In this thesis, the issue of construct underrepresentation arose primarily as an implication of constructing the L2 comprehensibility scale in Study 3. The scale was derived from a relatively small sample of L2 speakers and raters on a single task and has not yet been validated on an independent sample of speakers and raters using more varied

tasks or task types. In addition, the speakers were from a single L1 background, which limits the generalizability of the scale. It is likely, therefore, that some pertinent dimensions of the construct are not represented in the descriptors. For example, in Study 3, a small group of ESL teachers provided introspective reports on the linguistic factors that most strongly influenced their L2 comprehensibility judgments. However, a shortcoming of the think-aloud methodology is that introspective reports are only an indirect representation of what participants are actually thinking while carrying out a problem-solving (e.g., rating) task, and are incomplete accounts of their cognitive processes (see Ericsson & Simon, 1993; Green, 1998). Participants can only report on things they are conscious of and can "screen" what they choose to talk about or even provide misleading information.

In contrast to Study 2, which sought to elicit rater processes as they verbalized their thoughts while rating in real-time, the interest in the use of introspective techniques in Study 3 was to learn about raters' overall impressions of the L2 speech and influences on their scoring decisions, so that this data source could inform empirical rating scale development. To reflect this research interest and for practical reasons, teacher raters were not asked to verbalize their thoughts but, rather, typed their impressions into editable textboxes in a word processing document placed directly under each rating scale. Again, it is plausible that participants neglected to articulate certain criteria that were, in fact, relevant to their scoring decisions, due to not being consciously aware of these criteria, not knowing how to articulate them, not remembering them from their initial listening of the speech samples, not viewing them as important to the investigation, etc., with the result that these relevant dimensions are not mapped onto the developed scale. Clearly, a series of validation studies using different analytic techniques (e.g.,

multidimensional scaling, open-coding) needs to be carried out in further research to address the issue of the construct being too narrowly represented in the rating scale.

Construct Irrelevant-Variance

The second principal threat to test validity that Messick cites, construct-irrelevant variance, refers to factors that are extraneous to the construct being measured (Messick, 1990). Although measurement error is unrelated to the focal construct, Messick focuses his discussion on reliable variance, including excess variance associated with other distinct constructs, and test method variance. For example, an L2 oral pronunciation test could elicit extemporaneous speech samples using a prompt that some test-takers have repeatedly rehearsed that could unduly inflate their scores relative to ratings assigned to the unrehearsed productions of their peers. The variability that arises as a result of this practice effect for some but not all raters is likely to confound results of the measurement of L2 oral production and interfere with the trustworthiness of the inferences that are made on the basis of the test scores. Another scenario is that the content or format of picture elements in a picture narrative could make the task more difficult for some test-takers than others for reasons that do not have to do with the L2 ability being measured. If the task is found to systematically disadvantage test-takers from one linguistic or cultural background over another in a way that is not related to the ability being measured, then this would introduce bias into the assessment (see McNamara & Roever, 2006).

Messick (1998a) acknowledges that "what constitutes construct-irrelevant variance is a tricky and contentious issue" (p. 65), and a proviso to his original position of dismissing task factors as construct-irrelevant will be noted later in the discussion. As Upshur and Turner (1999) show, when it comes to objective, dichotomously scored items, the psychometric model is relatively simple: the test-taker interacts with the test task to

produce a score. In this scenario, the two sources of construct-irrelevant variance that Messick (1990) cites as detracting from the attribute being measured (i.e., the measure being too broad and method variance) need to be considered and minimized to the extent possible.

When the assessment is rater-mediated and the resulting scores are not objectively verifiable, the psychometric model becomes more complex and additional sources of variance need to be taken into account (Bachman et al., 1995; Upshur & Turner, 1999). One such factor (i.e., facet), rater characteristics, has the potential to influence both the quantitative scores that raters assign, and their qualitative approach to the rating task, including the criteria that raters attend to when making rating decisions (Brown et al., 2005; Eckes, 2008), and strategies to condense their complex impressions of an L2 performance into a numerical rating using the "superficial" rating descriptors (Cumming, Kantor, & Powers, 2001; Lumley, 2005).

Raters' perceptions of L2 speech as expressed through their scalar judgments and qualitative comments are main foci of this thesis. However, rater characteristics are not treated uniformly as interfering with measurement of the focal construct, as is described in the context of each paper below. ***Study 1*** examines individual differences in raters' musical ability/experience, phonological memory capacity, and attention control from a quantitative perspective. The goal was to examine whether these rater characteristics, which are not controlled for in real-world assessment situations (e.g., raters are not normally screened for musical ability), could unduly influence their scoring of L2 speech. Such variables are potential sources of construct-irrelevant variance and, if found to affect raters' score assignments, could jeopardize the inferences that are made on the basis of the ratings.

Study 2 examines experienced and novice raters' rating scale use and perceptions of the rating process, with ESL teaching experience as the criterion distinguishing experienced from novice raters. Notably, in Study 1, raters' cognitive abilities were treated as possible sources of construct-irrelevant variance. In contrast, rater experience in Study 2 could not simply be dismissed as noise in the data as or a threat to validity but, rather, was viewed as relevant to the broader issue of who should listen to and evaluate L2 speech. In L2 pronunciation research, for example, novice raters tend to be recruited more often than experienced raters for practical reasons (i.e., accessibility of the population). Arguably, eliciting ratings from both rater groups is ecologically valid, as these groups are likely to interact with beginner-level L2 speakers inside and outside of the classroom, respectively, although experienced raters are more likely to evaluate their speech in formal assessment contexts. Therefore, could novice raters' judgments serve as a "substitute" for those of experienced raters in research contexts on the basis of the numerical ratings they assign? And if experienced and novice raters are found to be equally consistent in their scoring and no different in the mean scores they assign, do they approach the rating task in the same way, use the same rating strategies, and attend to the same aspects of speech when making scoring decisions? Do novice raters experience more difficulties making scale level distinctions than experienced raters, particularly in mid-scale range? Study 2 constitutes a preliminary attempt to address these empirical questions. In sum, due to the nature of the inquiry, the variability associated with rater experience was not relegated to the status of a threat to validity in Study 2, but, rather, was regarded as a rich source of information that can reflect back on our understanding of the perceptual constructs under examination (Chalhoub-Deville, 1995).

Finally, building on the findings from Study 2, ***Study 3*** uses a mixed-methods approach to probe the criteria that feed into raters' L2 comprehensibility judgments for the purpose of empirical rating scale development. In contrast to Studies 1 and 2, between-group differences in rater background characteristics were not investigated in Study 3. Rather, raters' perceptions of their most salient scoring criteria and the quantitative linguistic measures most strongly associated with their mean L2 comprehensibility ratings were examined for utilitarian reasons—to learn more about influences on listeners' comprehensibility judgments so they could be distilled in rating scale descriptors. As with Study 2 and in contrast to Study 1, Study 3 treats rater variability as having the potential to reveal different dimensions of the construct. The study constitutes a preliminary attempt to understand the scope of construct-relevant criteria that could factor into raters' L2 comprehensibility judgments. However, as mentioned earlier, follow-up validation studies are needed to refine the scale, minimize construct underrepresentation, and to clarify the range of tasks and settings that scale descriptors can be extrapolated to (Brindley, 1998).

An additional source of construct-irrelevant variance that needs to be taken into account when the assessment process involves ratings, the characteristics of the rating scale, is subsumed under Messick's (1989) heading of "method variance." This refers to systematic effects associated with the measurement procedure that are extraneous to the construct of interest. Notably, the rating instrument that raters use to record their judgments is clearly distinct from another method effect used to obtain the scores—the task used to elicit the L2 performance (Upshur & Turner, 1999).

Whereas the rating scale property of the number of scale levels is manipulated in Study 2, task effects are not systematically examined in this thesis. In fact, the same

picture narrative task (the "suitcase narrative,") which has been used extensively in previous L2 speech research (e.g., Derwing, Munro, Thomson, & Rossiter, 2009), was used to elicit the L2 speech in all three studies. This task was selected because it has been shown to be effective in generating speech samples in adult L2 learners from a wide range of first language (L1) backgrounds and L2 proficiency levels, because it elicits extemporaneous speech samples while the picture prompt controls for content in a way that a personal narrative task would not, and, finally, because its extensive use in this line of research makes it ideal for use in a study that examines methodological research conventions (Study 2).

Although the task was not a direct focus of any of the studies presented here, the treatment of the task in relation to the definition of the construct surfaces in the construction of the empirically-derived L2 comprehensibility scale in Study 3, which is arguably the centerpiece of this thesis. However, before elaborating on the issue of the extent to which the task should be embedded in rating scale descriptors, it is important to first address the source of method variance that was the motivation for Studies 2 and 3—the rating scale. In the next section, a discussion of the perceived limitations in the way pronunciation is operationalized in existing rating scales is informed by insights from validity theory. This then leads to a discussion of the role of the task in the way constructs are defined and operationalized.

Challenges to Construct Definition and Operationalization and Applications for L2 Pronunciation

Comprehensibility, or listeners' perceptions of ease of understanding L2 speech (Derwing & Munro, 1997), is a major concept in L2 pronunciation research, is congruent with the instructional goal of helping learners achieve intelligible pronunciation, and is

central to interlocutors' communicative success in real-world interactions (Derwing & Munro, 2009b; Morley, 1994). Part of the impetus for developing the L2 comprehensibility scale in Study 3 is that the role of pronunciation within influential theoretical models of communicative competence (Canale & Swain, 1980) and communicative language ability (Bachman, 1990; Bachman & Palmer, 1996) is insufficiently delineated to inform rating scale development. It follows that there is little understanding of how pronunciation in general, and comprehensibility in particular, relate to the broader construct of L2 oral proficiency. This limitation is reflected in existing rating scales descriptors, which are characterized by the inconsistent treatment of pronunciation (e.g., ACTFL), by vague characterizations of comprehensibility (e.g., IELTS), or by conflating comprehensibility with accentedness (e.g., CEFR scale of Phonological Control) when they are partially independent dimensions (see Derwing & Munro, 2009b).

Challenges associated with defining and measuring constructs are also apparent in current L2 pronunciation research. For instance, the terms “intelligibility” and “comprehensibility” have not been consistently applied across studies, leading to definitional confusion and difficulties in making cross-study comparisons (Isaacs, 2008). Derwing and Munro (1997) provide useful and conceptually clear definitions of these terms, which have been adopted by several researchers (e.g., Kennedy & Trofimovich, 2008) and are also adhered to in this thesis. Intelligibility denotes the amount of the L2 learner’s utterance that the listener is able to understand, and is most often measured by the proportion of words that the listener is accurately able to transcribe, although other methods, such as true/false sentences (Derwing, Munro, & Wiebe, 1997) and comprehension questions (Hahn, 2004) have also been used. In contrast,

comprehensibility is defined as listeners' *perceptions* of ease of understanding of L2 speech and is most often measured on a numerical (Likert-type) scale punctuated with the scalar endpoints "extremely easy to understand" and "extremely difficult to understand" (Munro & Derwing, 1999).

As was noted above, the theoretical basis for understanding major constructs in L2 pronunciation research is poor. It follows that the chief distinction between intelligibility and comprehensibility, which both have to do with listeners' understanding of L2 speech, lies in the way that these concepts have been *operationalized*. That is, intelligibility, in most published studies, involves listeners' written transcriptions of the speech, whereas comprehensibility involves listeners' impressionistic ratings of the speech (Derwing & Munro, 1997).

Borsboom (2005), who examines measurement and validity theory from a psychological perspective, is critical of such an operational approach to construct definition. He argues that when constructs are solely operationally defined, they do not exist independently from the measurement apparatus. Comprehensibility, for example, necessitates a rating scale if listeners' impressions of ease of understanding are to be quantified, in the same way that the measurement of temperature presupposes a thermometer. Borsboom concludes that latent variable models (i.e., item response theory models and factor analysis methods) are the most tenable current psychometric models. In contrast to the operationalist position, in which the construct being measured is inextricably bound to the measurement procedure, in the realist position instantiated by the use of latent class models, the construct has the "existential status" of being distinct from the measurement procedure from which it is derived (p. 135). Borsboom presents evidence for this view by examining the statistical assumptions of the different

psychometric models, analyzing the "semantics" of their relevant mathematical formulae, and drawing on concrete examples, mostly from psychological tests, to demonstrate the ontological stances of the competing models (for a review of Borsboom's work, see Leighton, 2008).

In contrast to Messick's (1989) unitary view of validity as encompassing value judgments and social consequences, Borsboom views validity more restrictively as simply a property of a test. He argues that "a test is valid for measuring an attribute if and only if (a) the attribute exists, and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure" (p. 150). Clearly, an operationalist view does not constitute sufficient evidence that the attribute (construct) exists, since it cannot be teased apart from the method for measuring it. An approach Borsboom advocates that is aligned with his causation-based notion of validity is to first formulate a theory that describes the focal construct, in order to establish a strong substantive basis, and then to test the theory empirically by applying different analytic techniques, including psychometric models. One example Borsboom cites as a way of accomplishing this is to start off with a theory that postulates developmental stages (e.g., Pienemann's Processability Theory, 2005) and then "translate" this theory into a latent class model that, crucially, moves beyond correlational evidence to establish a *causal* link between the attribute being measured and the test-taker's response behavior.

Borsboom's (2005) proposal for providing evidence that a certain test is a valid measure of a given attribute may be viewed by some as the gold standard for psychological research. However, this approach cannot feasibly be implemented in L2 pronunciation research, even if it is considered desirable, due to an underdeveloped theoretical basis. In fact, the lack of a comprehensive theory for perceptual measures of

L2 pronunciation (e.g., comprehensibility) makes it difficult to separate the measurement instrument from the attribute itself, since there is no comprehensive theory to fall back upon. However, an empirical approach to understanding more about constructs is currently feasible and can eventually lead to the development of theory.

In this section of the overarching literature review, the attempt has been to apply Borsboom's critique of the instrumental approach to measurement to the way major holistic constructs in L2 pronunciation research have been defined. Feasible alternatives to an operational approach to construct definition are arguably limited, due to the absence of falsifiable theories. However, Borsboom's book is intended primarily for an audience of psychologists, and three essential points need to be made. The first is that, from the perspective of an L2 pronunciation researcher, it is clear that the volume of research on L2 pronunciation is much more limited than the psychological research that Borsboom refers to in his book. Similarly, more theoretical models have been proposed for psychological constructs such as "intelligence" than for L2 pronunciation-based constructs such as "intelligibility." On this basis, applying Borsboom's work to L2 pronunciation research may not seem appropriate, but could be useful for generating discussion and reflecting on current research and assessment practice. Second, constructs such as "intelligibility" have been defined in almost as many different ways as there are researchers (Isaacs, 2008; Rajadurai, 2007). Derwing and Munro (1997) have done a service to the field by creating a definitional distinction between intelligibility and comprehensibility that is conceptually simple, and it is the way these constructs are *operationalized* that makes this distinction clear. From this perspective, operationalizing constructs has led to some clarity in distinguishing between closely-related constructs, although a stronger substantive basis is arguably desirable. Falsifiable theories (possibly

that stem from empirical data) that can be rigorously empirically tested would be a sign that the field is maturing.

The third and final point that needs to be emphasized is that not all validity theorists or psychometricians adopt the same viewpoint as Borsboom (2005) regarding an instrumental approach to construct definition. In his 1994 article, Messick does not appear to place a value judgment when discussing his observation that constructs, such as general proficiency, are often articulated solely through the scoring rubrics of a rating instrument rather than being defined a priori. This is the case for the majority of L2 oral proficiency scales that are referred to in the literature review in Study 3. Messick elaborates that this practice leads to a task-centered approach rather than a construct-centered approach to performance assessment. The influence of the task on the rating descriptors may not always come through clearly in the descriptors, however.

Clearly, Messick's approach to construct validity is strongly construct-centered (i.e., where the starting point of test development is to specify the attribute and knowledge components that will be assessed) rather than task-centered (i.e., where the starting point of test development is to specify the task that will be assessed; see McNamara, 1996). While task-related variance was designated as construct-irrelevant without qualification in Messick's earliest publications describing his unitary concept of validity (e.g., 1989, 1990), Messick acknowledges in his 1994 paper that the notion of the task as a threat to validity in a task-centered approach is essentially meaningless, since the skills that are identified are relevant for task completion. Messick suggests that the central problem of the task-centered approach is that it can lead to a proliferation of task-specific rating scales that can solely be applied in the specific context in which they were developed. That is, limitations of the task-centered approach are replicability and

generalizability. Conversely, Messick suggests that adhering exclusively to a construct-centered approach risks generating scoring rubrics that are too generic, particularly if the construct "can legitimately have multiple manifestations at different levels of performance quality" (p. 17). To compensate for these weaknesses, Messick recommends finding a happy medium between task-centered and construct-centered approaches to rating scale development but offers little concrete guidance on how to strike this balance.

Chapelle's (1998) position in the L2 assessment literature could perhaps be considered middle of the road in that she advocates that construct definitions need to "specify relevant aspects of both trait and context" (p. 43). Although an in-depth discussion of Chapelle's "interactionalist perspective" is beyond the scope of this review, the essence of this approach to construct definition is that constructs are not simply the sum of the trait and the context. Rather, traits cannot be defined in isolation of the context, and contextual factors cannot be defined without reference to changes in the L2 learners' underlying ability. That is, the implication of defining the trait and ability in relation to one another is that they both undergo a change in quality. Chapelle uses interlanguage vocabulary as an illustrative example of this interactionalist approach to construct definition—a perspective that differs from the relegation of the task to the status of construct-irrelevant variance in Messick's initial unnuanced position (1990).

Divergent views on construct definition in the L2 assessment literature (see Bachman, 2007, for a review) are also reflected in different positions on the extent to which the language ability and the task should be intertwined in rating scales (Brindley, 1998; Fulcher, 2003; North, 2000; Turner & Upshur, 2002). This issue of the extent to which the task is embedded in the rating scale needed to be dealt with in Study 3. The scale was developed for the purpose of operationalizing the construct with more

precision, in addition to clarifying the criteria that are most salient for making level distinctions. The researchers' stance at the outset of rating scale development admittedly tended more toward the construct-centered side of the spectrum rather than the task-centered side of the spectrum, and this is reflected in task selection. The task conditions in the picture narrative do not directly resemble target language tasks. That is, telling a story spontaneously based on a picture prompt is not likely to be encountered very often for most adult L2 learners in their daily professional tasks (i.e., is relatively inauthentic), although it does relate more generally to their ability to spontaneously describe a sequence of events or to tell a coherent story in an L2. Thus, only a weak case could be made that the task and construct should be inextricably intertwined. On the other hand, the lexical items that the L2 learners produced are a concrete manifestation that the task does affect linguistic output.

The L2 speech samples in Study 3 were analyzed using segmental, suprasegmental, temporal, grammatical, lexical, and discourse-level measures, for a total of 19 measures. The discourse-level measures (propositions and cohesive devices) were examined because the picture narrative elicited discourse-level productions and could not have been examined had word- or sentence-level productions been elicited instead (Kennedy, 2009). Because it was not feasible to include all analyzed measures in the rating scale descriptors, concrete criteria for inclusion needed to be established based on the analyses.

If the discourse-level measures are ultimately included in the scale, then this would be a concrete acknowledgment of the influence of the task on the L2 productions. On the other hand, their exclusion would not necessarily imply that the task did not have bearing on the resulting L2 productions and would warrant further exploration (e.g., it

could have to do with the sensitivity of the measures). In sum, in this data-driven approach to rating scale development, the "solution" of whether to integrate the task into the rating scale descriptors was driven by the data. It is the researchers' intention to extensively pilot and refine the developed scale so it can eventually be used in the L2 classroom for formative-assessment purposes (i.e., as a tool for teaching and learning; see Colby-Kelly & Turner, 2007; Rea-Dickins, 2001).

Final Introductory Thought

The papers that make up the thesis are part of a research program that aims to better understand major constructs in L2 pronunciation research, improve current measurement practice, and, ultimately, reinvigorate the conversation on L2 pronunciation assessment, which has been virtually absent from the research agenda since the time of Lado (1961). Clearly, the views of validity theorists are central to any discussion on construct validity. However, it is important not to lose sight of the fact that listeners are by far the best resource for better understanding holistic constructs of L2 comprehensibility, accentedness, and fluency. Indeed, these constructs are defined in terms of listener *perceptions* (scalar judgments) of L2 speech. Thus, examining listeners' interpretations of the construct, listening and rating processes and strategies, and the alignment of their perceptions with linguistic characteristics of learner productions is essential to developing a greater understanding of the language ability that we are attempting to measure.

Introduction to Study 1

Language is viewed as a complex cognitive skill from an information processing perspective (O'Malley & Chamot, 1990). Variability in L2 attainment attributed to individual differences in L2 learner characteristics has long been of interest to SLA and psycholinguistic researchers. This is evidenced by a large volume of research on the topic, including books or edited volumes (e.g., Dörnyei, 2005; Robinson, 2002; Skehan, 1989) and numerous articles and book chapters (e.g., Dewaele, 2009; Ellis, 2004). A subset of this research invokes individual differences in L2 learner *cognitive abilities*, including language aptitude or analytic ability (Ranta, 2002; Sparks & Ganshow, 2001), phonological memory (French & O'Brien, 2008; O'Brien, Segalowitz, Collentine, & Freed, 2006), attention control (Segalowitz & Frenkiel-Fishman, 2005; Trofimovich et al., 2007a), and musical ability (Slevc & Miyake, 2006). These cognitive variables are investigated not only to explain differences in L2 learning outcomes, but also because they are presumed to relate to the processes underlying L2 development (Segalowitz, 1997).

There is also an emerging body of L2 assessment research that examines individual differences in rater *background characteristics*, although this research is relatively recent and less extensive. Some sources of rater variability that have been investigated include rater experience or expertise (Barkaoui, 2010; Cumming, 1990), linguistic background (Brown, 1995; Kim, 2009), gender (O'Loughlin, 2002, 2007), and orientations to different assessment criteria (Eckes, 2008; Turner, 2000). In addition, interlocutor effects have been examined in speaking tests, including oral interview formats (e.g., Brown, 2003, 2005), and paired and group oral tests (e.g., Davis, 2009; Van Moere, 2006).

One set of individual difference variables that has been examined in relation to L2 learner oral proficiency in the SLA/psycholinguistic literature but not in relation to raters' scalar judgments of L2 speech in the language assessment literature is *rater cognitive abilities*. Bringing together a psycholinguistic focus on cognitive variables with an assessment focus on rater background characteristics and scoring behavior, Study 1 investigates the influence of raters' musical ability, phonological memory capacity, and attention control on their scalar judgments of L2 speech.

Chapter 2 – Study 1

Phonological Memory, Attention Control, and Musical Ability: Effects of Individual Differences on Rater Judgments of L2 Speech

Talia Isaacs, McGill University

Pavel Trofimovich, Concordia University

Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of L2 speech. *Applied Psycholinguistics*, *32*, 113–140.

Phonological Memory, Attention Control, and Musical Ability:

Effects of Individual Differences on Rater Judgments of L2 Speech

Abstract

This study examines how listener judgments of second language (L2) speech relate to individual differences in listeners' phonological memory, attention control, and musical ability. Sixty native English listeners (30 music majors, 30 non-music majors) rated 40 non-native speech samples for accentedness, comprehensibility, and fluency. The listeners were additionally assessed for phonological memory (serial recognition), attention control (trail making), and musical aptitude. Results showed that music majors assigned significantly lower scores than non-music majors solely for accentedness, particularly for low ability L2 speakers. However, the ratings were not significantly affected by individual differences in listeners' phonological memory and attention control, which implies that these factors do not bias listeners' subjective judgments of speech. Implications for psycholinguistic research and for high-stakes speaking assessments are discussed.

Phonological Memory, Attention Control, and Musical Ability:

Effects of Individual Differences on Rater Judgments of L2 Speech

Introduction

As universities and other postsecondary institutions seek to attract an increasingly diverse student body, they face the responsibility of providing valid assessments of incoming students' language ability, especially when the students' mother tongue is not the language of instruction (Cheng, Myles, & Curtis, 2004). There have been attempts to develop technology-based, automated assessment instruments for spoken English (Downey, Farhady, Present-Thomas, Suzuki, & Van Moere, 2008). However, the most commonly used second language (L2) speaking tests in academic settings, whether they rely on recorded speaking prompts (e.g., TOEFL iBT, TSE) or on face-to-face interaction (e.g., IELTS), are scored by human raters (Templer, 2004). Rater judgments in academic settings are central to high stakes decisions, including whether or not a candidate is admitted to the university, placed in a remedial language course, or awarded a teaching assistantship.

Although rater judgments are often used as the chief source of evidence of L2 speakers' language proficiency in academic settings, such judgments might not always be reliable (e.g., when scoring is not internally consistent) or valid (e.g., when scoring is influenced by factors extraneous to the construct being measured). That is, raters' judgments might reflect not simply speakers' performance, but also individual differences among raters themselves. For example, ongoing validation research of standardized L2 tests such as the TOEFL, TSE, and IELTS has revealed various sources of rater

variability (Brown, Iwashita, & McNamara, 2005; Myford & Wolfe, 2000; Taylor, 2007), including raters' experience (Cumming, 1990), gender (O'Loughlin, 2007), the relative weight they place on different scoring criteria (Eckes, 2008), and their native language (L1) background (Kim, 2009). What is not known, however, is whether other sources of rater variability, for example, those related to individual differences in raters' *cognitive abilities* (e.g., phonological memory, attention control, or music aptitude) also influence raters' assessments of spoken language.

In the present study, we therefore investigated whether individual differences in raters' phonological memory (auditory working memory capacity), attention control (ability to allocate attention efficiently), or musical skill (musical aptitude) influence raters' judgments of L2 speech on dimensions of accentedness, comprehensibility, and fluency. Accentedness is defined here as listeners' judgments of how closely the pronunciation of an utterance approaches that of a native speaker (Munro & Derwing, 1999). Comprehensibility refers to listeners' perceptions of how easily they understand an utterance (Munro & Derwing, 1999). Fluency denotes listeners' assessments of how smoothly and rapidly an utterance is spoken (cf. Derwing, Rossiter, Munro, & Thomson, 2004). Our overall goal was to determine how phonological memory, attention control, and musical ability could contribute to listeners' perceptual judgments of L2 speech and, as a result, could influence their scoring decisions.

Phonological Memory

Phonological memory (also referred to as phonological short-term memory) refers to a language user's capacity to retain spoken sequences temporarily in a short-term memory store. This capacity is usually associated with the phonological loop, a subcomponent of the human working memory system responsible for temporary storage

of verbal-acoustic information (Baddeley, 2003; Baddeley & Hitch, 1974). Often measured in terms of language users' ability to recall digits or repeat nonwords, phonological memory is a strong predictor of vocabulary knowledge in both L1 and L2 (French & O'Brien, 2008; Gathercole, Hitch, Service, & Martin, 1997; Masoura & Gathercole, 2005). Other evidence has implicated phonological memory in the development of L2 grammar (Ellis & Sinclair 1996; O'Brien, Segalowitz, Collentine, & Freed, 2006; Trofimovich, Ammar, & Gatbonton, 2007) and L2 speaking (Fortkamp, 1999; O'Brien, Segalowitz, Freed, & Collentine, 2007). Phonological memory is also a predictor of overall L2 learning success, as assessed through classroom grades or standardized tests (Kormos & Sáfár, 2008). In this study, we hypothesized that phonological memory plays a role in listeners' perceptual judgments of L2 accentedness, comprehensibility, and fluency.

The link between phonological memory and speech perception is well established. Early experiments showed that listeners perceive speech in a speech-specific manner, relying on phonological memory to do so. For example, Baddeley, Lewis, and Vallar (1984) examined the phonological similarity effect. These researchers showed that listeners recall phonologically dissimilar items better than similar ones. This finding suggests that speech is encoded in a temporary phonological memory store, where similar sounding items are subject to considerably more interference and are thus harder to recall than dissimilar items. In another line of research, Rowe and Rowe (1976) studied the so-called stimulus suffix effect (see also Morton, Crowder, & Prussin, 1971). These researchers had listeners recall sequences composed of either speech (words) or non-speech (environmental sounds). Each sequence was followed by an extraneous "suffix", which was also either speech (e.g., the word *go*) or non-speech (e.g., a bird chirp). The

results extended a finding from previous research that the presence of a suffix impairs listeners' recall (Conrad, 1960; Crowder & Morton, 1969), and suggested that this disruption occurs when the suffix matches the type of sequence to be recalled (speech or non-speech). This implies that listeners tend to rely on different mechanisms to process speech versus non-speech material, with phonological memory involved in the processing of speech and an acoustic storage system involved in the processing of non-speech (Crowder & Morton).

Recent evidence points to a more direct role of phonological memory in speech perception. For example, phonological memory is involved in listeners' ability to discriminate stress contrasts not present in their L1 (Dupoux, Peperkamp, & Sebastián-Gallés, 2001) and in listeners' perceptual learning of words, especially when such words are degraded to make the learning task more difficult (Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008). In addition, phonological memory seems to underlie, at least in part, listeners' ability to perceive spoken sentences, as shown in tasks requiring listeners to detect mispronunciation or to comprehend sentences involving minimal pairs (Jacquemot, Dupoux, Decouche, & Bachoud-Lévi, 2006). Phonological memory also appears to be related to listeners' subjective ratings of speech. For instance, Gould, Saum, and Belter (2002) reported a relationship between listeners' recall of spoken directions, their phonological (working) memory, and their subjective reactions to speech (e.g., rating the speaker as being kind and caring vs. patronizing and disrespectful).

In light of these findings, we hypothesized that listeners' perceptual judgments of L2 speech might also be related to phonological memory. However, there is little in existing literature to indicate how phonological memory could influence ratings of speech in general, and ratings of L2 speech in particular, that would favor one set of predictions

over another, allowing us to propose a directional hypothesis. For example, listeners with a large phonological memory, as compared to listeners with a small memory capacity, could retain more L2 speech in their short-term memory, which would make the task of listening to L2 speech easier for them. As a result, they could judge L2 speech as being more comprehensible, more fluent, and less accented. On the other hand, listeners with a large phonological memory, precisely because of their superior memory capacity, could be highly sensitive to various phonetic and prosodic deviations of L2 speech from L1 "norms" and, consequently, could judge L2 speech as being less comprehensible, less fluent, and more accented. We investigated these possibilities here by studying the link between individual differences in listeners' phonological memory and their perceptual judgments of L2 speech.

Attention Control

Attention control refers to an individual's ability to efficiently allocate attention among different aspects of language or different cognitive processing tasks. As a cognitive construct, attention control involves a number of functions associated with a variety of neurobiological structures (Posner & DiGirolamo, 2000). When applied to language, attention control may refer to enhanced processing of the linguistic stimuli that are relevant to the task at hand and to inhibited processing of the stimuli that are irrelevant to it (Eviatar, 1998). Attention control may also refer to an individual's ability to shift attention efficiently among different sets of linguistic relationships (Talmy, 1996).

The existing literature on attention control and speech perception is extensive (see Cowan & Saults, 1995, and Cowan et al., 2005), dating back to early studies of selective attention (e.g., Cherry, 1953) and its conceptualizations in theories of information processing (e.g., Atkinson & Shiffrin, 1968). Several findings from this literature are

pertinent here. One finding is that attention control is implicated in speech processing at all levels, from fine-grained phonetic perception to higher-order semantic processing. At the level of phonetic perception, for example, efficient attention control might be required for listeners to perceive phonetic cues signalling voicing distinctions (Gordon, Eberhardt, & Rueckl, 1993) and vowel contrasts (Assmann & Summerfield, 1994), especially when listeners perform multiple tasks at once. And at the level of speech comprehension, recall of speech is often disrupted when listeners' attention is divided, suggesting that speech comprehension draws on substantial attentional resources (Craik & McDowd, 1987; Wood & Cowan, 1995).

Several other findings from attention literature suggest that speech perception tasks require efficient attention control, especially when such tasks are performed under non-ideal listening conditions. One example of such tasks includes monitoring speech for a particular speech segment (e.g., /b/ or /v/) when listening to two competing spoken messages (Mullennix, Sawusch, & Garrison, 1992). Another example is listening to speech and detecting errors in it while performing a secondary (concurrent) task (Oomen & Postma, 2002). It is likely that these tasks might be comparable (at least to a certain extent) in their demands on the listener to the task of listening to L2 speech, particularly if L2 speech is accented, difficult to understand, and dysfluent. To understand L2 speech, listeners may need to allocate their attention efficiently to several competing dimensions in speech, including its phonetic and perceptual aspects and its semantic content (von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003), particularly when these dimensions do not correspond to what listeners might consider nativelike in their language.

It is conceivable, then, that listeners' judgments of accentedness, comprehensibility, and fluency of L2 speech might be related to individual differences in

listeners' attention control. In other words, listeners' ability to optimize a balance in attentional focus between processing the interlocutor's speech on the one hand and constructing a mental representation of the message on the other might underlie, at least in part, their ability to process and evaluate L2 speech. Similarly, listeners may also need to switch their attention seamlessly among different linguistic dimensions when attending to speech, and these "shift costs" may vary depending on how heavily accented, difficult to understand, or dysfluent L2 speech seems to listeners.

Based on previous research, we were again unable to predict the exact nature of the relationship between listeners' judgments of L2 speech and their attention control capacity. On the one hand, listeners with efficient attention control could shift their attentional focus effortlessly among different linguistic dimensions of speech or between the tasks of constructing perceptual and conceptual representations of speech. As a result, these listeners could judge L2 speech as being more comprehensible, more fluent, and less accented than listeners with less efficient attention control. On the other hand, listeners with efficient attention control could be overly sensitive to additional shift costs imposed on them by L2 speech because of some extra effort needed for parsing the linguistic dimensions of speech or for constructing its conceptual representation. Compared to listeners with less efficient attention control, these listeners could downgrade their ratings of L2 speech and could judge it to be less comprehensible, less fluent, and more accented. We explored these possibilities here by investigating the link between individual differences in listeners' attention control and their perceptual judgments of L2 speech.

Musical Ability

We hypothesized that individual differences in musical ability might also be

related to how listeners evaluate L2 speech. For the purposes of this study, musical ability is defined as an individual's ability to "hear" (internalize) music that is no longer present in the physical environment, a skill that Gordon (1995) termed "audiation". For example, upon hearing two musical phrases played consecutively, listeners with greater musical ability, as compared to listeners with weaker musical ability, would presumably be able to judge whether the two phrases are similar in their melodic contour (overall pattern of pitch rises and falls), even if the two phrases differed in the overall number of notes. Musical ability, defined in this manner, is often measured using standardized tests which target several aspects of this ability, including pitch, intensity, rhythm, timbre, tonal memory, and timing (Bentley, 1966; Gordon, 1995; Seashore, 1919; Wing, 1968).

It appears that musicians (i.e., individuals who are presumably good at the skills involved in how we have defined musical ability) are at an advantage over non-musicians in a variety of speech perception tasks. For example, in a series of behavioral and neuroimaging experiments, Schön, Magne, and Besson (2004) demonstrated that musicians are more accurate than non-musicians at detecting melodic incongruities (tones that violate musical or prosodic contours) in both music and L1 speech. Similarly, Gottfried (2007) showed that trained musicians are more adept than non-musicians at perceiving (and producing) the lexical tones of an unfamiliar tone language (Mandarin), with musicians outperforming non-musicians in both tone discrimination tasks and goodness-of-production ratings assigned by native speaking listeners. Alexander, Wong, and Bradlow (2005), who reported a similar finding in Mandarin pitch perception tasks, in fact suggested that musicians' extensive pitch processing experience may positively transfer to speech perception.

Although these results point to an important link between musical ability and L1 speech processing, the relationship between musical ability and L2 speech processing remains unclear. Some researchers who have investigated this relationship reported a positive correlation between musical ability and L2 production (Arellano & Draper, 1972; Nakata, 2002; Slevc & Miyake, 2006). However, many others have failed to reveal any clear relationship between these two variables (Dexter & Omwake, 1934; Flege, Munro, & MacKay, 1995; Pimsleur, Stockwell, & Comrey, 1962; Tahta, Wood, & Loewenthal, 1981). The link between musical ability and L2 speech perception has been even more elusive, essentially because this relationship has been studied much less extensively. For example, Slevc and Miyake (2006) showed that a standardized measure of musical ability accounted for up to 12% of variance in native Japanese speakers' perception of L2 (English) contrasts in words, sentences, and spoken texts (see also Pimsleur et al., 1962). However, no association between musical ability and L2 perception was found in several other studies (Arellano & Draper, 1972; Nakata, 2002).

Clearly, more research is needed not only to enhance our understanding of the link between musical ability and L2 processing, but also to explore the interface between musical ability and the assessment of L2 speech. It is not clear, for example, whether trained musicians would judge L2 speech differently than listeners with little or no musical experience and with lower musical ability. In this study, we therefore investigated the link between individual differences in listeners' musical ability and their perceptual judgments of L2 speech. Based on previous research (e.g., Alexander et al., 2004; Gottfried, 2007), we predicted that listeners with greater musical ability will judge L2 speech as being more accented, less comprehensible, and less fluent than raters with weaker musical ability. We reasoned that listeners with greater musical ability would be

more sensitive to certain aural components of L2 speech (e.g., non-native voice quality or pitch fluctuations) than listeners with weaker musical ability (see Gottfried, 2007). As a result, those listeners who are more musical, compared to those who are less musical, would have a lower impression of the L2 speech they heard, assigning lower scores for accentedness, comprehensibility, and fluency.

The Current Study

To the best of our knowledge, the relationship between cognitive variables, which underlie any form of language functioning, and listener judgments of L2 speech has not been examined in prior research. Therefore, the overall goal of this study was to investigate the extent to which native speaking listeners' judgments of L2 speech are mediated by individual differences in listeners' phonological memory, attention control, and musical ability. To accomplish this goal, we asked 60 listeners (half of whom were formally trained musicians and half not) to rate the speech of 40 francophone L2 speakers of English for accentedness, comprehensibility, and fluency. We also asked all listeners to perform three cognitive tasks: a serial non-word recognition task to measure listeners' phonological memory, the Trail Making Test to measure their attention control, and three subtests of the Musical Aptitude Profile to measure their musical ability. We then analyzed the speech ratings as a function of these three cognitive measures to determine how the speech ratings related to the cognitive measures.

Method

Speakers

The speech samples for this study were elicited from 40 adult native speakers of French (27 female, 13 male) tested as part of a larger, unrelated project (Trofimovich, Gatbonton, & Segalowitz, 2007b). All speakers (mean age: 35.6 years, range: 18–61)

were born into francophone families in Québec and were educated in French. With the exception of two, whose first exposure to English occurred between birth and age two through interaction with an English-speaking parent, all speakers were first exposed to English in elementary school (mean age: 9.3 years) as part of English as a second language instruction in Québec. Prior to providing speech samples, the speakers rated their proficiency in speaking and listening in English and French on a 9-point scale (1 = *extremely poor*, 9 = *extremely proficient*) and estimated their daily use of French and English on a 0-100% scale. In French, the mean ratings for the two skills were consistently high (8.9–9.0); in English, they were intermediate (5.7–6.6) and highly variable. On average, the speakers used French 80% (30–100%) and English 20% (0–70%) of the time daily.

As part of the earlier project (Trofimovich et al., 2007), a separate test was administered to determine that the speakers represented a wide range of L2 pronunciation ability. The test was a reading task in which the speakers read a simple 440-word story in English and were recorded directly onto a computer using a Plantronics (DSP-300) microphone. The recordings were subsequently presented to a panel of five judges (mean age: 38.2 years; all exposed to English from birth) to assess the degree of accentedness in the speakers' speech. The judges listened to a short excerpt from each speaker's recording (mean duration: 18 s) and independently rated each sample for accentedness using a 9-point Likert-type scale (1 = *heavily accented*, 9 = *not accented at all*). An accent score was computed for each speaker by averaging the five judges' accent ratings (interrater reliability: $\alpha = .96$). The scores ranged between 1.8 and 9.0, with a mean of 5.3. The

speakers thus represented different pronunciation ability levels, from beginning to advanced.

For the current study, we used samples of extemporaneous speech recorded by the 40 speakers in response to a simple eight-frame picture narrative. Used in previous studies with L2 speakers (e.g., Derwing et al., 2004), the picture narrative depicted a man and a woman who, having bumped into one another on a busy street corner, realized their mishap of having switched suitcases only after they had arrived at their respective destinations. The speakers were asked to study the picture narrative for approximately one minute prior to recording their story directly onto a computer (using a Plantronics DSP-300 microphone). The recorded stories ranged in duration between 26.4 and 322.8 seconds. Excerpts containing the first 20 seconds of each story, excluding initial pauses and false starts, were then saved separately as digital audio files, normalized for peak intensity, and randomized for their presentation to raters. The procedure of having raters judge the first few seconds of a speech sample (cf. Derwing, Thomson, & Munro, 2006) has the advantage of keeping the content of the story relatively constant across speakers in a naturalistic, extemporaneous speech task, where the precise output of the speaker is unpredictable.

Raters

The raters who listened to and evaluated the speech samples included 60 native English-speaking undergraduate students (26 males, 34 females). Of these, 30 were music majors enrolled in a music program at an English-medium university in Montreal, Canada, and 30 were non-music majors studying a variety of disciplines at the same university (e.g., psychology, political science, electrical engineering, English literature, computer science). All raters were native speakers of English from either the United

States (31) or English speaking areas of Canada (29), with a similar proportion of Americans in the music and non-music major groups (53% and 47% respectively). One music major reported being a monolingual English speaker, 24 in both rater groups cited knowledge of an L2, and 6 music and 7 non-music majors indicated knowing a third language as well. Overall, 41 raters reported that their L2 was French, eight cited Spanish, four identified German, and the rest named other minority languages. All raters reported having normal hearing and none had had language teaching experience or had taken a phonetics/phonology course, although six of the voice majors had taken a diction course for singers. A summary of the raters' background information, which includes their estimates of language use and their proficiency self-ratings in French, appears in Table 1.

The music majors consisted of 19 performance majors, 6 music education students, 4 Bachelor of Arts music majors (music theory or music history concentrations), and a composition major (mean self-reported musical experience: 10.5 years, range: 3–19 years). The primary instruments for the music majors included string instruments (8), voice (7), woodwinds (7), brass (3), keyboard (3), and percussion instruments (2). In addition, the majority (24) had received formal training in a second or third instrument. By the time of the testing, all music majors had completed a minimum of one year of required courses in musicianship (ear training) and music theory (tonal counterpoint and harmony analysis). Although three of the performance majors were jazz musicians, they had received training in the Western classical music tradition during their first year of university coursework.

Table 1. *Raters' Background and Language Proficiency Characteristics*

Measure	Music majors		Non-music majors	
	M	*SD*	*M*	*SD*
Chronological age[a]	20.6	1.5	20.8	1.8
Residence in Montreal[a]	2.8	3.8	2.9	1.5
Age of L2 learning[a]	10.8	3.3	11.8	4.4
French listening self-rating[b]	4.2	2.2	3.3	1.9
French speaking self-rating[b]	3.4	1.9	2.7	1.7
English use in listening/speaking[c]	90.3	7.6	93.3	8.6
French use in listening/speaking[c]	9.8	7.7	5.8	7.4
Exposure to nonnative speakers[c]	38.5	18.2	34.7	13.6

Note: [a]In years. [b]Measured on a 9-point scale (1 = *extremely poor*, 9 = *extremely proficient*). [c]Measured on a 0-100% scale.

The non-music majors, who were not pursuing a university music degree and had no aspiration to become professional musicians, had varying degrees of musical training (mean self-reported musical experience: 3.4 years, range: 0–9 years), with eight reporting no musical training at all. Our intent was not to recruit a homogeneous group of non-musicians with absolutely no musical background. Instead, we chose to access a group who, along with the music majors, varied in the quantity (and likely also the quality) of musical experience they had received. Obtaining a wide distribution of musical ability was consistent with our goal here, namely, to determine whether listeners' musical ability and other cognitive variables influence their accentedness, comprehensibility, and fluency ratings.

A series of independent samples *t*-tests was conducted to determine if the two rater groups differed for any of the eight demographic and language use variables listed in Table 1. These tests revealed no statistically significant differences between the groups, suggesting that both rater groups were matched on all demographic and language use variables examined here.

Phonological Memory Task

A serial non-word recognition task was used in this study to measure phonological memory. In this task, listeners hear sequences of pronounceable non-words (with sequences increasing in length as the task progresses) and decide whether the sequences are presented in the same or in a different order. A recognition task of this kind has at least two advantages over widely-used measures of phonological memory based on word/non-word recall and repetition. First, a recognition task does not contain an articulatory (motor) component. In contrast, both recall and repetition tasks place articulatory demands on the speaker and likely show bias against participants with speech impediments (Snowling, Chiat, & Hulme, 1991). Second, a recognition task, compared to recall and repetition tasks, appears to minimize lexical (vocabulary knowledge) influences on phonological memory, yielding a relatively accurate estimate of phonological memory (Gathercole, Pickering, Hall, & Peaker 2001). For example, Gathercole et al. reported partial η^2 (effect size) values of .71–.94 for lexical influences in a serial recall task as compared to effect sizes of .25–.27 for lexical influences in a serial recognition task.

The materials for the serial recognition task used here consisted of 160 one-syllable CVC non-words from Gathercole et al. (2001). The non-words (which respected

English phonotactics) were digitally recorded by a male native English speaker and were organized into sequences of five, six, and seven items, with eight pairs of sequences of each length (for a total of 24 pairs). All items within a sequence had a different vowel sound, and the consonant composition within each sequence was as distinctive as possible. Half of the pairs were ordered identically (e.g., *loog jahl deech kerp meb … loog jahl deech kerp meb*, where ellipses indicate a short pause). In the other half of the pairs, one of the items was transposed in the second presentation of the sequence relative to the first, so that there was an order mismatch (e.g., *lod tudge jick norb garm … lod tudge norb jick garm*). To reduce the salience of transposed items and to encourage the listener to process the complete sequence, the first or the last item in the sequence was never transposed. The location of the transposed items was varied randomly across sequence lengths.

The 24 non-word sequences were presented to the raters over a Koss R/80 headset using speech presentation software (Smith, 1997). The five-item sequence pairs were presented first, followed by the six- and seven-item sequence pairs. The non-words were presented at the rate of one item every 800 ms, with a 1.5 s pause between the two sequences in each pair. Upon hearing each pair, the raters indicated whether the two sequences were presented in the same or a different order by clicking one of the buttons labeled “same order” and “different order” on the computer screen. The raters had unlimited time to provide their judgment but were not permitted to replay the sequence or to change their response. Prior to carrying out this task, the raters were given two same and two different sequence pairs as practice. The number of correct responses (out of 24) was recorded for each rater and used as a measure of phonological memory.

Attention Control Task

The Trail Making Test, originally designed as part of the U.S. Army Individual Test Battery (1944), was used in this study to estimate attention control. The test appears to provide a language-neutral estimate of an individual's ability to shift attention between two sets of stimuli (Lee, Cheung, Chan, & Chan, 2000). The test consists of two parts and involves drawing a line to connect consecutive digits from 1 to 25 (1-2-3-4-5-6, etc.) in Part A and drawing a similar line to connect alternating digits and letters (1-A-2-B-3-C, etc.) in Part B. Assuming that the time it takes to complete a non-alternating digit sequence (Part A) provides the baseline for each individual's motor and visual control, the additional cost imposed on the individual by the alternating digit-letter sequence (Part B) provides a measure of this individual's executive control, or the ability to switch attention between two stimulus sequences. In other words, the difference in completion time between Part B and Part A of the test is indicative of the individual's attentional control of switching between different stimuli (Corrigan & Hinkeldey, 1987) and between different cognitive tasks (Arbuthnott & Frank, 2000).

For all raters, Part B of the test followed Part A, each preceded by an eight-item practice session. The completion times for both parts of the test were measured using a digital stopwatch and were recorded in seconds, with the values rounded to the nearest one hundredth of a second. For each rater, the difference in completion times between Part B and Part A of the test was used as a measure of attention control. A smaller score, corresponding to a smaller difference in completion time between Parts A and B, represented more efficient attention control.

Test of Musical Ability

Three subsections (melody, tempo, phrasing) of the Musical Aptitude Profile

(MAP), a test battery used to predict musical learning in grade four to college-level students (Gordon, 1967, 1995, 2001; see also Carson, 1998), were used to measure musical ability. The test assumes no prior knowledge of music other than general exposure to music. A lack of musical training does not preclude an individual from receiving a high score on the test, although individuals who do have musical training are more likely to receive high scores than those with no musical training (Gordon, 2001).

In the melody subtest, listeners hear two consecutive short musical phrases performed on a violin and judge whether the first musical phrase (stimulus item) sounds similar to or different from the second musical phrase (test item) in terms of melodic contour (overall pattern of pitch rises and falls). In the tempo subtest, listeners hear two musical phrases and judge tempo consistency. If the tempo is inconsistent, the test item gradually speeds up or slows down relative to the tempo that had been established in the stimulus. Listeners are required to indicate whether the tempo in the two musical excerpts is the same or different. Finally, in the phrasing subset, which was performed by violin and cello, listeners hear the same musical phrase twice and judge which rendition they feel sounds better in terms of phrasing (i.e., is performed more musically). The intent of this subtest is to go beyond tonal and rhythmic dimensions of music to assess interpretive aptitude by measuring listeners' responses to the combined effect of dynamics, tempo, tone quality, and musical articulation (see Gordon, 1995, for details on the validation of MAP subtests). We used these three subtests because they covered a range of skills that could potentially differentiate among individuals with stronger and weaker musical ability.

There were no forced-choice responses for any of the subtests, and listeners had the option of indicating "unsure" (counted as a non-response in the scoring) if they did

not want to venture a guess. The number of correct responses was calculated for each rater on each subtest and used as measures of musical ability. The melody and tempo subtests were scored out of 40, and the phrasing subtest was scored out of 30.

Procedure

The testing, which took approximately two hours to complete, was conducted individually in a quiet room using a desktop computer and a Koss R/80 headset. The raters first listened to the 40 speech samples presented one at a time in one of four randomized orders and rated each sample for accentedness, comprehensibility, and fluency using separate numerical scales. These three constructs were operationalized based on previous L2 research (e.g., Derwing et al., 2004; Kennedy & Trofimovich, 2008; Munro & Derwing, 1999): accentedness (1 = *heavily accented*, 9 = *not accented at all*), comprehensibility (1 = *hard to understand*, 9 = *easy to understand*), and fluency (1 = *not fluent at all*, 9 = *very fluent*). The listening session was self-paced, and the raters were allowed to listen to each recording, re-play its segments, and change their responses as many times as they wished. With rare exceptions, all maintained an efficient scoring pace, making rating decisions without frequent re-playing of recordings and changing of the ratings given. The raters then completed the three subsections of the MAP. All MAP instructions and musical excerpts were played on the CDs included with the test battery, and the raters marked their responses on the standardized scoring sheets. Finally, the raters performed the serial non-word recognition test and the Trail Making Test (in that order).

Results

For all statistical tests, the alpha level for significance was set at .05. A Bonferroni procedure was applied to adjust the level of significance for all multiple comparisons. All

t-tests and correlations are based on two-tailed distributions. Effect sizes are reported as *r*.

Preliminary Analyses

Prior to examining the relationship between cognitive variables and ratings of accentedness, comprehensibility, and fluency of L2 speech, we performed four preliminary analyses. The goal of the first analysis was to determine the relationship among the three cognitive measures. We computed Pearson correlation coefficients among the raters' (n = 60) phonological memory scores, attention control scores, and their scores on the three MAP subtests (α = .005). The three music scores were significantly correlated with one another, $r(58)$ = .43–.69, $p < .001$, suggesting that the three subtests measured a related construct. By contrast, neither the phonological memory score nor the attention control score was significantly correlated with each other or with the music scores, $r(58)$ = -.13–.09, $p > .32$, suggesting that the three cognitive measures focused on here represented separate constructs.

The goal of our second analysis was to determine whether there were any differences between the two rater groups (music and non-music majors) for the three cognitive measures investigated here. We computed independent samples *t*-tests comparing the two groups (α = .01). These tests yielded significant differences between the groups for all three MAP subtests: melody, $t(58) = 5.67$, $p < .0001$, $r = .60$, tempo, $t(58) = 3.79$, $p < .0001$, $r = .45$, and phrasing, $t(58) = 2.75$, $p = .01$, $r = .34$. In all cases, the music majors outperformed the non-music majors. By contrast, these tests yielded no significant differences between the two groups for phonological memory and attention control ($p > .92$). Thus, the music and non-music majors, as intended, differed only in their musical ability. Mean scores for the music and non-music majors on the MAP

subtests, serial non-word recognition task, and Trail Making Test are provided in Table 2.

Table 2. *Mean Scores on Cognitive Ability Tests*

Measure	Music majors		Non-music majors	
	M	*SD*	*M*	*SD*
MAP melody subtest[a]	38.03	1.61	32.07	5.43
MAP tempo subtest[a]	38.80	1.47	35.70	4.29
MAP phrasing subtest[b]	23.03	3.93	20.33	4.21
Serial non-word recognition tasks[c]	16.00	3.07	15.93	2.86
Trail Making Test[d]	8.58	9.54	8.52	8.32

Note: [a]Scored out of 40. [b]Scored out of 30. [c]Scored out of 24. [d]Calculated as the difference in completion time between Part B and Part A (seconds).

In the third analysis, we focused on rater reliability. This was done to determine whether a given rater's behavior was consistent with the other raters in their group. We assessed rater reliability separately for each rater group (the 30 music majors and 30 non-music majors) for each of the three ratings. To allow for comparison of reliability coefficients across studies, we computed two interrater reliability measures: Cronbach's alpha and mean intercorrelations (Pearson *r*) corrected for distortion using Fisher's *Z* transformation. The obtained Cronbach's alpha values for both groups ranged between .98 and .99 for the three ratings. These values were about .03 to .04 higher than those reported for untrained novice raters in Derwing and Munro (1997) and Derwing et al. (2004), where rater sample sizes were 26 and 28 respectively, and comprehensibility,

accentedness, and fluency were operationalized using 9-point numerical rating scales. The obtained mean Pearson's *r* values for both rater groups were .76 for the three ratings. These values were about .05 higher than the values obtained by Munro and Derwing (2001), who examined 48 raters' judgments of speakers from multiple L1 groups, but were nearly equivalent to the values reported by Derwing et al. (2004), who tested speakers from a homogenous L1 background. Thus, the interrater reliability for both rater groups here was sufficiently high for listeners with no rater training. Based on these analyses, we computed a single mean accentedness, comprehensibility, and fluency score for each rater by averaging across each rater's 40 individual accentedness, comprehensibility, and fluency ratings, respectively.

In our final analysis, we closely examined the distribution of these mean accentedness, comprehensibility, and fluency scores to determine if they were appropriate for parametric analyses. We conducted a series of Kolmogorov-Smirnov goodness-of-fit tests, separately for the music and non-music majors, to compare whether the distributions of these mean scores were different from normally distributed sets of scores with the same means and standard deviations. These tests yielded no significant values for any ratings, Ds(30) < .12, p > .20, suggesting that the assumption of normality was met and that the mean rating scores could be analyzed using parametric procedures.

Phonological Memory and Ratings of L2 Speech

The raters' phonological memory scores ranged between 10 and 23, with a mean of 15.9 and a median of 16. To determine the relationship between phonological memory and ratings of comprehensibility, accentedness, and fluency, we divided the entire sample of raters (n = 60) into two equal groups using a median split: those whose phonological memory scores were above the median value (mean: 18.2, range: 16–23) and those whose

memory scores were below this value (mean: 13.6, range: 10–15). We then examined whether there were differences between these two groups in their mean ratings of accentedness, comprehensibility, and fluency (shown for both groups in Table 3). These comparisons ($\alpha = .016$) yielded no statistically significant differences between the groups of high and low phonological memory, ts(58) < .53, ps > .60, rs < .07. This finding suggested that, at least in this study, there was no relationship between the raters' phonological memory and their ratings of L2 speech.

Table 3. *Mean Accentedness, Comprehensibility, and Fluency Ratings (Standard Deviations) as a Function of Phonological Memory (Low, High), Attention Control (Worse, Better), and Musical Training (Non-Music Majors, Music Majors)*

Rating	Phonological Memory		Attention Control		Musical Training	
	Low	High	Worse	Better	Non-music majors	Music majors
Accentedness	5.18 (1.12)	5.03 (1.01)	4.94 (1.07)	5.26 (1.04)	5.41 (1.03)	4.79 (1.01)
Comprehensibility	6.23 (1.34)	6.24 (0.99)	6.12 (1.09)	6.36 (1.23)	6.47 (1.12)	6.00 (1.17)
Fluency	5.36 (1.05)	5.31 (0.78)	5.20 (0.84)	5.48 (0.97)	5.54 (0.84)	5.14 (0.95)

Attention Control and Ratings of L2 Speech

The raters' attention control scores ranged between -9.7 (for a rater who was actually faster in Part B than in Part A of the test) and 31.7, with a mean of 8.6 and a median of 8.3. As in the previous analysis, to determine the relationship between attention control and ratings of comprehensibility, accentedness, and fluency, we used a median

split to divide the entire sample of raters (n = 60) into a group of raters with better attention control, that is, those whose scores fell below the median value (mean: 1.3, range: -9.7–8.3) and a group of raters with worse attention control (mean: 15.8, range: 8.4–31.7). As before, we tested for differences between these two groups in their mean ratings of accentedness, comprehensibility, and fluency (shown for both groups in Table 3). These comparisons (α = .016) yielded no statistically significant differences between the groups of better and worse attention control, ts(58) < 1.21, ps > .23, rs < .16. This finding suggested that, at least in this study, there was no relationship between the raters' attention control and their ratings of L2 speech.

Musical Training and Ratings of L2 Speech

Because in our preliminary analyses we determined that our original groups of the music and non-music majors differed significantly in their musical ability, as measured by the three MAP subtests, we proceeded to examine whether these two groups differed in their mean ratings of accentedness, comprehensibility, and fluency (shown in Table 3). These comparisons (α = .016) yielded a difference only for accentedness, t(58) = 2.37, p = .021, r = .30, with the music majors assigning significantly lower mean ratings than the non-music majors. Although the music majors tended to score both comprehensibility and fluency more negatively than the non-music majors (see Table 3), neither of these differences was statistically significant, ts(58) < 1.73, ps > .09, rs < .22. These results suggest that there might be a difference in how the music and non-music majors rate accentedness in L2 speech (although this difference, at p = .021, failed to reach statistical significance after a Bonferroni adjustment). We explored this finding in greater detail in a follow-up analysis.[1]

In comparisons between the groups of the music and non-music majors, we used speech ratings that were averaged across the 40 speakers. However, these mean ratings conceal much variability. At least some of this variability is specific to differences in speakers' ability and to differences in how negatively listeners rate speakers of different ability. One possible way to explore the relationship between musical ability and accentedness ratings further is to examine the music and non-music majors' ratings of accentedness for speakers of different ability. In order to accomplish this, we divided our original sample of 40 speakers into separate groups based on the accentedness ratings (1 = *heavily accented*, 9 = *not accented at all*) given to these speakers by an independent group of raters as part of an earlier study (Trofimovich et al., 2007). These ratings (described in the *Speakers* section above) allowed us to create three groups: heavily accented speakers (n = 15, mean accentedness rating: 2.9, range: 1.8–3.8), speakers with intermediate accent (n = 13, mean: 5.5, range: 4.2–6.4), and unaccented speakers (n = 12, mean: 8.0, range: 7.0–9.0).

For each of the 60 raters, we computed three mean accentedness ratings by averaging each rater's accentedness ratings across the speakers in each group. These mean accentedness ratings, which were normally distributed according to Kolmogorov-Smirnov tests, Ds(30) < .13, p > .21, were then submitted to comparisons between the groups of the music and non-music majors. These tests (α = .016) yielded a significant difference in accentedness ratings between the music and non-music majors for the heavily accented speakers, t(58) = 2.61, p = .011, r = .32. However, the difference between the two groups for the speakers with intermediate accent became nonsignificant after a Bonferroni correction was applied, t(58) = 2.23, p = .03, r = .28. In both of these

cases, the music majors assigned lower accentedness scores (i.e., rated them as sounding less nativelike) than the non-music majors. Finally, no difference was detected between the two groups for the unaccented speakers, $t(58) = 1.52$, $p = .13$, $r = .19$. It appears, then, that raters with university-level musical training assigned lower accentedness ratings than raters with little or no musical experience, especially when rating L2 speakers of lower ability. The mean accentedness ratings given by the raters to the speakers of different ability are plotted in Figure 1.

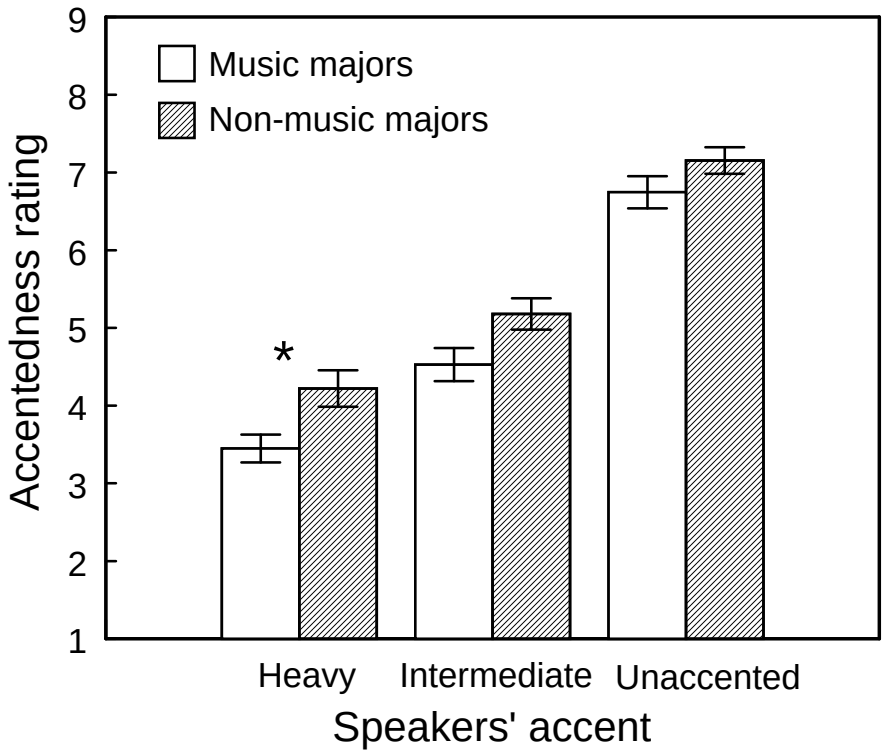

Figure 1. Mean accentedness ratings by music and non-music majors for L2 speakers with heavy, intermediate, and little accent. Brackets enclose ± 1 SE. An asterisk designates a significant difference between the two rater groups after a Bonferroni adjustment.

In a subsequent analysis, we explored further the relationship among the three ratings for the music and for the non-music majors. Our intention here was to determine how ratings of accentedness, comprehensibility, and fluency relate to one another for individuals with musical training and those with no academic training in music. For this analysis, we first computed three Pearson product moment correlation coefficients among the mean ratings of accentedness, comprehensibility, and fluency, separately for the music and non-music majors. These correlation coefficients (shown in Table 4) revealed that all correlations were overall lower for the music majors than for the non-music majors. Next, we applied the Fisher *r*-to-*z* transformation (Clark-Carter, 1997) to compare the correlations obtained from the music and non-music majors (independent samples). The correlation between accentedness and comprehensibility was significantly higher for the non-music majors than for the music majors ($z = -1.97$, $p < .05$). The correlation between accentedness and fluency was also higher for the non-music majors than for the music majors ($z = -1.93$, $p = .053$); this difference, however, narrowly missed statistical significance. This suggests that the three rating dimensions, and accentedness and comprehensibility in particular, were more independent (distinct) for the music majors than for the non-music majors.

Table 4. *Pearson Product-Moment Correlations Among Accentedness, Comprehensibility, and Fluency Ratings for Music Versus Non-Music Majors*

Measures	1	2	3
	Music majors		
1 Accentedness			
2 Comprehensibility	.47**		
3 Fluency	.58**	.81**	
	Non-music majors		
1 Accentedness			
2 Comprehensibility	.78**		
3 Fluency	.83**	.80**	

* $p < .005$, ** $p < .001$, two-tailed.

In our final analysis, we sought to examine the relationship between accentedness, comprehensibility, and fluency for raters classified into different musical ability groups based on their performance on the MAP subtests. A measure of musical ability derived from a standardized test is arguably a more objective criterion for classifying raters into musical ability groups than their university major status, which at least partially reflects raters' self-selection into a program with music specialization. In order to examine musical ability in relation to the rating outcomes, therefore, we split the raters into three musical ability groups based on their composite scores on the MAP subtests, irrespective of their university major status, with an equal number of raters placed into each group.

The raters assigned to the low ability group received a mean MAP composite score of 82.35 (range: 51–92), the intermediate ability group's mean composite score was 96.45 (range: 93–99), and the high ability group achieved a mean score of 103.15 (range: 100–108). In light of the finding reported above that accentedness appears to be more distinct (i.e., linearly independent) for the music majors than for the non-music majors relative to comprehensibility and fluency, we predicted lower correlation coefficients between each of the three rating dimensions (and in particular between accentedness and comprehensibility) for higher musical ability raters than for lower musical ability raters.

The Pearson correlation coefficients for the three musical ability groups are reported in Table 5. Results of significance tests using the Fisher *r*-to-*z* transformation showed that the correlation between accentedness and comprehensibility was significantly higher for the low ability group than for the high ability group ($z = -2.33$, $p = .020$) and approached significance for the low versus the intermediate ability groups ($z = -1.90$, $p = .057$). In addition, the correlation between accentedness and fluency was significantly higher for the low ability group than for both the intermediate ($z = -2.03$, $p = .004$) and the high ($z = -2.12$, $p = .003$) ability groups. As in the previous analysis, there were no significant differences among the groups in the correlation coefficients obtained for comprehensibility and fluency. To summarize, when raters were grouped into high, intermediate, and low musical ability groups based on their composite MAP scores, raters with higher musical ability appeared to differentiate accentedness from comprehensibility or fluency to a greater extent than raters with lower musical ability. These findings are virtually identical to those we reported in the previous analysis, where we compared raters with university musical training to those with no university musical training.

Table 5. *Pearson Product-Moment Correlations Among Accentedness, Comprehensibility, and Fluency Ratings for Low, Intermediate, and High Musical Ability Raters Grouped by the Composite Score on the MAP Subtests*

Measures	1	2	3
High musical ability			
1 Accentedness			
2 Comprehensibility	.29		
3 Fluency	.43	.75**	
Intermediate musical ability			
1 Accentedness			
2 Comprehensibility	.74**		
3 Fluency	.82**	.90**	
Low musical ability			
1 Accentedness			
2 Comprehensibility	.80**		
3 Fluency	.83**	.76**	

* $p < .005$, ** $p < .001$, two-tailed.

Discussion

Our overall motivation in conducting the present study was to determine the role of cognitive variables in rater judgments of L2 speech, and to understand how these cognitive variables affect the reliability of language assessments and ultimately influence

high-stakes decision making. We started from the premise that English language speaking proficiency in North American higher educational institutions is typically assessed through such standardized tests as the TOEFL, IELTS, or TSE, all used for admission or placement purposes or for the selection of international teaching assistants (Fox, 2005; Luoma, 2001). Although there has been some research into sources of rater variability (e.g., Cumming, 1990; Eckes, 2008; Kim, 2009), ours appears to be the first study that examines how individual differences in raters' cognitive abilities impact their judgments of L2 speech. To this end, we analyzed accentedness, comprehensibility, and fluency ratings of L2 speech as a function of raters' phonological memory, attention control, and musical ability.

There were two main findings. Our first main finding was that the speech ratings examined here did not depend on listeners' phonological memory or attention control. This finding suggests that individual differences in raters' phonological memory and attention control (at least insofar as they were measured here) do not play a strong role in rater judgments of accentedness, comprehensibility, and fluency. This result is reassuring because these potential biasing effects, which are not relevant to the rating constructs, do not seem to threaten the validity of the speaking assessments scored by human raters. Our second main finding was that the speech ratings examined here depended on raters' musical training, such that university-trained musicians tended to assign lower mean scores than musically untrained raters. These differences were the most pronounced when raters evaluated the accentedness of L2 speech, especially for speakers of low pronunciation ability. Accentedness and comprehensibility also appeared to be more independent (distinct) dimensions for university trained musicians than for musically untrained raters. This result suggests that musical training, which was strongly associated

with musical ability in this study, is a factor that could bias L2 speaking assessments. We discuss both of these main findings in turn.

Phonological Memory, Attention Control and Assessments of Speaking

Our first main finding was that raters' judgments of L2 speech did not depend on individual differences in raters' phonological memory and attention control. From an assessment point of view, this finding is encouraging as it suggests that individual differences in these two cognitive abilities might not contribute to unwanted variance in speech ratings. From a psycholinguistic perspective, however, this finding raises interesting questions regarding the precise contribution of phonological memory and attention control to listener judgments of speech. To the best of our knowledge, our study was the first to begin addressing these questions.

In setting up our study, we argued that phonological memory, given its extensive involvement in a variety of speech processing tasks (Baddeley, 2003), could be involved in listener judgments of L2 speech. There are at least two reasons why we found no evidence for this. The first reason relates to the measure of phonological memory used here. We employed a serial non-word recognition task to estimate the raters' phonological memory capacity. It could be that other tasks (e.g., non-word repetition or recall tasks), despite their shortcomings that we attempted to sidestep here (see Gathercole et al., 2001), could yield a measure of phonological memory that would be associated with perceptual judgments of L2 speech. Another (and perhaps more plausible) reason is that perceptual judgments of speech may not draw heavily on phonological memory. This is because phonological memory, as its name suggests, operates on phonological, not necessarily rich physical (acoustic-phonetic) details of the speech signal (Baddeley, 2003). If listeners rely on physical details of speech (such as pitch fluctuations or

phonetic substitutions) for their ratings of L2 speech, then it is not surprising that individual differences in *phonological* memory do not appear to influence these ratings. Perhaps what could play a role in perceptual ratings of L2 speech is *acoustic* memory, which refers to an individual's capacity for storing acoustic-phonetic information in speech (Cowan, 1984). Research on the suffix effect (discussed earlier) would seem to support the interpretation that the raters in this study were primarily storing and retaining acoustic information in the short-term store (Crowder & Morton, 1969; Rowe & Rowe, 1976). That is, it appears that the rating task, which entailed listening to but not verbally recalling the speech, likely encouraged the raters to encode the speech as a series of sounds or syllables (i.e., using bottom-up processes) rather than as words or conceptual categories (cf. Bloom, 2006). Since acoustic memory is involved in a variety of language processing tasks (Cowan & Saults, 1995), it would be interesting to explore its contribution to perceptual judgments of speech.

At the outset of this study, we also predicted that attention control could be involved in listener judgments of L2 speech. We defined attention control broadly, as an individual's ability to efficiently allocate attention among different aspects of language (e.g., separate linguistic dimensions of speech) or different cognitive processing tasks (e.g., constructing perceptual and conceptual representations of speech). Given the extensive evidence showing the involvement of executive attention control in speech processing tasks (Cowan & Saults, 1995; Cowan et al., 2005), our failure to find a significant association between attention control and perceptual ratings of speech could be an artefact of our testing procedure. It is likely that the listeners in this study did not need to rely extensively on their attention control capacity, simply because the task of rating L2 speech, as implemented here, was not cognitively demanding and therefore did

not require listeners to exercise efficient attention control. This interpretation is supported by the results of one previous study which used the Trail Making Test to estimate participants' attention control. In that study, the measure of attention control was a stronger predictor of participants' performance when the cognitive demands of the task were elevated (Trofimovich et al., 2007). Thus, it would appear that providing perceptual judgments of L2 speech may not be a cognitively demanding task for a native speaker of a language, and perhaps even judging L2 speech that is highly accented, difficult to understand, and dysfluent does not call extensively on listeners' attentional resources.

In a recent overview of attention and its role in various cognitive tasks, Cowan and his colleagues (Cowan et al., 2005) offered yet another reason for why attention control (as it was conceptualized here) might not be relevant to perceptual judgments of speech. These researchers suggested that a more meaningful measure of attention and its role in processing tasks should be the *scope* of attention, and not necessarily its control. Broadly, the scope of attention, as defined by Cowan et al. (2005), refers to "the capacity of the focus of attention" (p. 49). The scope of attention is assumed to be specific to an individual language user, and its size is believed to be related to a language user's working memory capacity and intellectual aptitude. In future investigations of the role of cognitive factors in perceptual judgments of L2 speech, it would be interesting to examine these claims further. Such an investigation could, for example, employ a test of the scope of attention (Cowan, Fristoe, Elliott, Brunner, & Saults, 2006) to examine whether and to what extent the scope of attention can be predictive of listeners' perceptual judgments of speech.

Musical Training and Assessments of Speaking

Our second main finding was that raters' judgments of L2 speech depended on raters' musical training, which was strongly associated in this study with raters' musical ability. This finding shows an important link between musical training and L2 speech processing, and adds to a growing body of research that reveals cognitive consequences of musical training and aptitude for the perception and production of L2 speech (e.g., Arellano & Draper, 1972; Pimsleur et al., 1962; Slevc & Miyake, 2006). However, unlike the findings of previous studies which show positive effects of musical ability on perception and production, our findings point to a potentially negative, biasing effect of musical training on native speaking listeners' judgments of L2 speech.

It is important to bear in mind that, in the present study, musical training appeared to be associated statistically significantly only with raters' judgments of L2 accentedness, although a similar association (albeit a weak one) was also found for judgments of L2 comprehensibility and fluency. This raises an interesting question of how essential accentedness ratings are to speaking assessment of L2 speakers. Previous research has shown that accentedness, as it was defined in this study, tends to be associated with lower ratings than either comprehensibility or fluency (e.g., Derwing & Munro, 1997; Munro & Derwing, 1999). One of the most robust findings from this body of research is that a strong non-native accent does not necessarily impede intelligibility (extent to which listeners understand L2 speech) although unintelligible speech is almost always judged to be heavily accented (Derwing & Munro, 2005). If the goal of language teachers, assessment specialists, and L2 learners themselves is for learners to be fully intelligible in their L2, as opposed to sounding like a native speaker (Levis, 2005), then perceptual ratings of speech, especially accentedness ratings, should not be done in isolation but

should always be tied to an assessment of how well L2 speakers are understood (see Jenkins, 2000, for a similar argument). In other words, accentedness judgments, when carried out in the absence of assessment of L2 speakers' intelligibility, can be misleading and biasing, and ultimately not particularly useful.

Although ratings of L2 accentedness done in isolation might not be particularly revealing of overall L2 speaking ability, it is nevertheless important to understand precisely why musically trained raters appear to assign lower scores than musically untrained raters. Accentedness ratings have been shown to correlate with prosodic aspects of L2 speech (e.g., intonation, pitch accent) for speakers of several languages (Anderson-Hsieh & Koehler, 1988; Mareüil & Vieru-Dimulescu, 2006). Given a growing body of evidence for music-language transfer effects at the prosodic level (e.g., Patel, Peretz, Tramo, & Labreque, 1998), it is likely that musicians' enhanced sensitivity to certain aspects of L2 speech, particularly at the level of prosody, is linked to their lower accentedness ratings. Future research could attempt to isolate those accent-related aspects of L2 speech that lend themselves to differences between musically trained and less musically experienced raters.

In future research, it would also seem appropriate to investigate the precise contribution of musical training and experience to the reliability and validity of L2 speaking assessment. In this study, the musically trained raters tended to judge L2 speech more negatively than the raters with little or no musical training (although the accentedness variable was the only one that reached statistical significance). At the same time, however, accentedness appeared to be a more independent (distinct) dimension relative to comprehensibility and fluency for music majors and raters with higher musical ability than for non-music majors and raters with lower musical ability. Thus, although

musical training might lead to L2 speech ratings which could be more biased toward the negative end of the rating scale, these ratings might more precisely target each of the constructs being measured. This raises an intriguing possibility which needs to be investigated in future research: namely, that musical training (i.e., experience through which listeners get implicitly sensitized to certain aspects of L2 speech) and rater training (i.e., explicit instructions and practice about certain aspects of L2 speech given to raters prior to assessment) have a similar impact on the rater. Both types of experiences might sensitize raters to those aspects of L2 speech that are relevant to each construct being measured, ultimately leading to more accurate assessment of L2 speech.

Implications

Although interesting from a research perspective, our findings may not at this time have immediate implications for real-world high-stakes assessments, where standards for reliability need to be high and sources of rater variability should be minimized to the extent possible in the interests of test fairness. In high-stakes assessments, raters are also calibrated in a norming process, with the overall goal of minimizing individual raters' scoring idiosyncrasies in order to achieve greater homogeneity of scoring. Clearly, the current study was not conceived with such a high-stakes assessment context in mind. Therefore, it remains unclear whether and to what extent raters' musical experience (or musical ability) would reduce the interrater reliability or pose a threat to the validity of raters' subjective judgments of L2 speech in a high-stakes assessment setting. It may be that raters with more musical experience assign a greater weighting to a particular aspect of speech (e.g., segmental errors that contribute to the impression of an L2 accent) than raters with less musical experience, whether or not those aspects of speech are specified in the rating scale descriptors (e.g., Eckes, 2008; Elder, Barkhuizen, Knoch, & von

Randow, 2007). It may also be that musical experience contributes to systematic differences in rater leniency or severity, a source of variability that researchers have sought to address with rater training (e.g., Elder, Knoch, Barkhuizen, & von Randow, 2005; Lumley & McNamara, 1995).

Even if it is established in future research focusing on operational L2 speaking tests that rater behavior differs as a function of musical experience, the implication is not necessarily that judgments of a musically homogenous group of raters should be sought. Rather, perhaps a rater training component could be introduced to help mitigate the musical experience effect. The scope of such an intervention could be determined, in part, by considering what aspects of speech musically experienced raters are overly sensitive to and whether those aspects are relevant to the construct being measured. This and other factors, such as the feasibility of the training and its real-world applicability, could be used to decide which raters should serve as the norming group. It would be interesting to examine the extent to which a rater training procedure would be effective in getting raters to attend to the target aspect of speech, and whether this procedure would alter their ratings in the desired direction. Such considerations, which lie beyond the scope of the present study, are fertile ground for future research.

Concluding Remarks

In the present study, we examined the role of cognitive variables in rater judgments of L2 speech, with an overall goal of understanding how cognitive variables could affect the validity of language assessments and could ultimately influence high-stakes decision making. However, this study is only a first attempt to address this research goal. It is still unclear how cognitive variables affect different types of speaking assessments, such as paired tasks or tasks specific to a profession or occupation, where

the construct of speaking ability is perhaps more broadly defined (e.g., incorporating grammar or vocabulary). Likewise unknown are the effects of many other individual differences, for example, field dependence or analytical ability, on the assessment of speaking. Musical ability, the factor that appeared to influence scalar judgments of L2 speech in this study, also needs to be explored in greater detail in order to understand precisely the nature of the impact of musical expertise and experience on rater behavior. These and other issues remain to be explored in future research.

Acknowledgments

This research was supported by a Social Sciences and Humanities Research Council of Canada (SSHRC) doctoral fellowship to the first author and by both SSHRC and Fonds québécois de la recherche sur la société et la culture (FQRSC) grants to the second author. The authors would like to thank Randall Halter for his invaluable statistical advice, Tracey Derwing, Murray Munro, Irena O'Brien, and Norman Segalowitz for sharing some of their testing materials, Sarita Kennedy for her assistance with participant recruitment, Hyojin Song for her help with participant selection and testing, and two anonymous AP reviewers for their insightful comments on earlier versions of this paper.

References

Alexander, J. A., Wong, P. C. M., & Bradlow, A. R. (2005). Lexical tone perception in musicians and non-musicians. In *Proceedings of Interspeech 2005, Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal. Retrieved March 11, 2009, from faculty.wcas.northwestern.edu/ann-bradlow/Alexander-Wong-Bradlow-2005.pdf

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning, 38*, 561-613.

Arbuthnott, K., & Frank, J. (2000). Trail Making Test, Part B as a measure of executive control: Validation using a set-switching paradigm. *Journal of Clinical and Experimental Neuropsychology, 22*, 518-528.

Arellano, S. I., & Draper, J. E. (1972). Relations between musical aptitudes and second-language learning. *Hispania, 55*, 111-121.

Assmann, P. F., & Summerfield, Q. (1994). The contribution of waveform interactions to the perception of concurrent vowels. *Journal of the Acoustical Society of America, 95,* 471-484.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89-195). New York: Academic Press.

Baddeley, A. D. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36*, 189-208.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advance in research and theory* (Vol. 8, pp. 47-89). New York: Academic Press.

Baddeley, A., Lewis, V., & Vallar, G. (1984). Exploring the articulatory loop. *Quarterly Journal of Experimental Psychology, 36A,* 233-252.

Bentley, A. (1966). *Measures of musical abilities*. London: Harrap.

Bloom, L. C. (2006). Two-component theory of the suffix effect: Contrary evidence. *Memory & Cognition, 34*, 648-667.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks*. TOEFL Monograph 29. Princeton, NJ: Educational Testing Service.

Carson, A. D. (1998). Why has musical aptitude assessment fallen flat? And what can we do about it? *Journal of Career Assessment, 6*, 311-328.

Cheng, L., Myles, J., & Curtis, A. (2004). Targeting language support for non-native English-speaking graduate students at a Canadian university. *TESL Canada Journal, 21*, 50-71.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, *25*, 975-979.

Clark-Carter, D. (1997). *Doing quantitative psychological research: From design to report*. Hove, UK: Psychology Press.

Conrad, R. (1960). Very brief delay of immediate recall. *Quarterly Journal of Experimental Psychology, 12*, 45-47.

Corrigan, J., & Hinkeldey, N. (1987). Relationships between Parts A and B of the Trail Making Test. *Journal of Clinical Psychology, 43*, 402-409.

Cowan, N. (1984). On short and long memory stores. *Psychological Bulletin, 96,* 341-370.

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., &

Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51,* 42-100.

Cowan, N., Fristoe, N. M., Elliott, E. M., Brunner, R. P., & Saults, J. S. (2006). Score of attention, control of attention, and intelligence in children and adults. *Memory and Cognition, 34,* 1754-1768.

Cowan, N., & Saults, J. S. (1995). Memory for speech. In H. Winitz (Ed.), *Human communication and its disorders, a review* (Vol. IV, pp. 83-170). Timonium, MD: York Press.

Craik, F. I. M., & McDowd, J. M. (1987). Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 474-479.

Crowder, R. G., & Morton, J. (1969). Precategorical acoustic storage. *Perception & Psychophysics, 5*, 365-373.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*, 31-51.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*, 1-16.

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

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*, 665-679.

Derwing, T. M., Thomson, R. I., & Munro, M. J. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System, 34*, 183-193.

Dexter, E. S., & Omwake, K. T. (1934). The relation between pitch discrimination and

accent in modern languages. *Journal of Applied Psychology, 18*. 267-271.

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Van Moere, A. (2008). Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly, 5*, 160-167.

Dupoux, E., Peperkamp, S., & Sebastian-Galles, N. (2001). A robust method to study stress "deafness". *Journal of the Acoustical Society of America, 110,* 1606-1618.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*, 155-185.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*, 37-64.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training. Does it work? *Language Assessment Quarterly, 2*, 175-196.

Ellis, N. C., & Sinclair, S. G. (1996). Working memory and the acquisition of vocabulary and syntax: Putting language in good order. *Quarterly Journal of Experimental Psychology, 49A*, 234-250.

Eviatar, Z. (1998). Attention as a psychological entity and its effects on language and communication. In B. Stemmer & H. A. Whitaker (Eds.), *Handbook of neurolinguistics* (pp. 275- 287). New York: Academic Press.

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

Fortkamp, M. B. M. (1999). Working memory capacity and elements of L2 speech

production. *Communication and Cognition*, *32*, 259-295.

Fox, J. (2005). Re-thinking second language (L2) admission requirements: Problems with language-residency criteria and the need for language assessment and support. *Language Assessment Quarterly, 2*, 85-115.

French, L. M., & O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics, 29,* 463-487.

Gathercole, S., Hitch, G. J., Service, E., & Martin, A. J. (1997). Phonological short-term memory and new word learning in children. *Developmental Psychology, 33*, 966-979.

Gathercole, S. E., Pickering, S. J., Hall, M., & Peaker, S. M. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *Quarterly Journal of Experimental Psychology, 54A*, 1-30.

Gordon, E. E. (1967). Implications for the use of the "Musical Aptitude Profile" with college and university freshman music students. *Journal of Research in Music Education, 15*, 32-40.

Gordon, E. E. (1995). *Manual: Musical Aptitude Profile*. Chicago: GIA Publications.

Gordon, E. E. (2001). *A three-year study of the Musical Aptitude Profile*. Chicago: GIA Publications.

Gordon, P. C., Eberhardt, J. L., & Rueckl, J. G. (1993). Attentional modulation of the phonetic significance of acoustic cues. *Cognitive Psychology, 25,* 1-42.

Gottfried, T. L. (2007). Music and language learning: Effect of musical training on learning L2 speech contrasts. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 221-237). Amsterdam: John Benjamins.

Gould, O. N., Saum, C., & Belter, J. (2002). Recall and subjective reactions to speaking styles: Does age matter? *Experimental Aging Research, 28,* 199-213.

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 460-474.

Jacquemot, C., Dupoux, E., Decouche, O., & Bachoud-Lévi, A.-C. (2006). Misperception in sentences but not in words: Speech perception and the phonological buffer. *Cognitive Neuropsychology*, *23*, 949-971.

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

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

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*, 187-217.

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition, 11,* 261-271.

Lee, T. M. C., Cheung, C. C. Y., Chan, J. K. P., & Chan, C. C. H. (2000). Trail making across languages. *Journal of Clinical and Experimental Neuropsychology, 22*, 772-778.

Levis, J. (2005). Pronunciation and the assessment of spoken language. In R. Hughes (Ed.), *Spoken English, applied linguistics and TESOL: Challenges for theory and*

practice (pp. 245-270). London: Palgrave Macmillan.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*, 54-71.

Luoma, S. (2001). A review of the Test of Spoken English (TSE). *Language Testing, 18*, 225-234.

Mareüil, P., & Vieru-Dimulescu, B. (2006). The contribution of prosody to the perception of foreign accent. *Phonetica, 63*, 247-267.

Masoura, E. V., & Gathercole, S. E. (2005). Contrasting contributions of phonological short-term memory and long-term knowledge to vocabulary learning in a foreign language. *Memory, 13*, 422-429.

Morton, J., Crowder, R. G., & Prussin, H. A. (1971). Experiments with the stimulus suffix effect. *Journal of Experimental Psychology, 91,* 169-190.

Mullennix, J. W., Sawusch, J. R., & Garrison, L. F. (1992). Automaticity and the detection of speech. *Memory and Cognition, 20,* 40-50.

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

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

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English assessment system*. TOEFL Research Report 65. Princeton, NJ: Educational Testing Service.

Nakata, H. (2002). Correlations between musical and Japanese phonetic aptitudes by native speakers of English. *Reading Working Papers in Linguistics, 6,* 1-23.

O'Brien, I., Segalowitz, N., Collentine, J., & Freed, B. (2006). Phonological memory and lexical, narrative, and grammatical skills in second-language oral production by adult learners. *Applied Psycholinguistics*, *27*, 377-402.

O'Brien, I., Segalowitz, N., Freed, B. F., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition, 29*, 557-582.

O'Loughlin, K. (2007). An investigation into the role of gender in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 63-97). Cambridge: Cambridge University Press.

Oomen, C. C. E., & Postma, A. (2002). Limitations in processing resources and speech monitoring. *Language and Cognitive Processes, 17,* 163-184.

Patel, A. D., Peretz, I., Tramo, M., & Labreque, R. (1998). Processing prosodic and musical patterns: A neuropsychological investigation. *Brain and Language, 61*, 123-144.

Pimsleur, P., Stockwell, R. P., & Comrey, A. L. (1962). Foreign language learning ability. *Journal of Educational Psychology, 53,* 15-26.

Posner, M. I., & DiGirolamo (2000). Attention in cognitive neuroscience: An overview. In M. S. Gazzaniga & E. Bizzi (Eds.), *The new cognitive neurosciences.* (pp. 623-631). Cambridge, MA: MIT Press.

Rowe, E. J., & Rowe, W. G. (1976). Stimulus suffix effects with speech and nonspeech sounds. *Memory and Cognition, 4,* 128-131.

Schön, D., Magne, C., & Besson, M. (2004). The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology, 41*,

341-349.

Seashore, C. E. (1919). *The psychology of musical talent*. New York: Silver, Burdett.

Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency: Does musical ability matter? *Psychological Science, 17,* 675-681.

Smith, S. C. (1997). *UAB Software*. Department of Rehabilitation Sciences: University of Alabama at Birmingham.

Snowling, M., Chiat, S., & Hulme, S. (1991). Words, nonwords, and phonological processes: Some comments on Gathercole, Willis, Emslie & Baddeley. *Applied Psycholinguistics, 12*, 369-373.

Tahta, S., Wood, M., & Loewenthal, K. (1981). Foreign accents: Factors related to transfer of accent from the first language to a second language. *Language and Speech*, *24*, 265-272.

Talmy, L. (1996). The windowing of attention. In M. Shibatani & S. A. Thompson (Eds.), *Grammatical constructions* (pp. 235-288). Oxford: Oxford University Press.

Taylor, L. (2007). The impact of the joint-funded research studies on the IELTS Speaking Module. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 185-194). Cambridge: Cambridge University Press.

Templer, B. (2004). High-stakes testing at high fees: Notes and queries on the international English proficiency assessment market [Electronic Version]. *Journal for Critical Education Policy Studies*, *2*. Retrieved March 11, 2009 from http://www.jceps.com/?pageID=article&articleID=21

Trofimovich, P., Ammar, A., & Gatbonton, E. (2007). How effective are recasts? The role of attention, memory, and analytical ability. In A. Mackey (Ed.), *Conversational*

interaction in second language acquisition (pp. 171-195). Oxford: Oxford University Press.

Trofimovich, P., Gatbonton, E., & Segalowitz, N. (2007). A dynamic look at L2 phonological learning: Seeking processing explanations for implicational phenomena. *Studies in Second Language Acquisition, 29*, 407-448.

US Army Individual Test Battery. (1944). *Manual of directions and scoring*. Washington, DC: Cambridge University Press.

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research, 17,* 48-55.

Wing, H. D. (1968). *Tests of musical ability and appreciation: An investigation into the measurement, distribution, and development of musical capacity* (2nd ed.). London: Cambridge University Press.

Wood, N., & Cowan, N. (1995). The cocktail party phenomenon revisited: Attention and memory in the classic selective listening procedure of Cherry (1953). *Journal of Experimental Psychology: General, 124,* 243-262.

Notes

1. Although we found that the music majors were significantly more negative than the non-music majors in their judgments of L2 accentedness, it was unclear whether this result would also hold when the raters' judgments of L2 accentedness were examined in relation to their musical *ability* (as opposed to their musical training or experience). To examine this issue, we divided the raters by median split into high and low musical ability groups using their MAP composite scores. An independent-samples *t*-test revealed no significant difference in L2 accentedness ratings between the two groups, although the low musical ability group overall tended to assign higher (i.e., more nativelike) accentedness ratings (5.26) than the high musical ability group (4.94). Therefore, in the discussion section, we discuss our findings only in relation to the raters' musical training and musical experience, not their musical aptitude. Clearly, the precise relationship between listeners' musical ability (aptitude), musical training, and their evaluative reactions to speech warrants closer inspection in future research.

Connecting Text – Study1 to Study 2

Study 1 examined the influence of individual differences in raters' musical ability, phonological memory capacity, and attention control on their ratings of L2 comprehensibility, accentedness, and fluency. Overall, these variables exerted little influence on the mean ratings of the L2 speech, as was evident in two main findings. First, no significant effects were detected for phonological memory and attention control, which removes these variables as possible sources of rater bias. Second, music majors assigned lower mean ratings than non-music majors solely for accentedness. However, this finding was relatively weak, in that the effect was not robust to a Bonferroni correction ($p > .016$) and yielded only a small effect size ($r = .3$).

Beyond issues of statistical significance, it is also unclear how *practically* significant the accentedness finding is. Although accents are perpetually salient, most L2 pronunciation researchers do not regard accent reduction as an appropriate goal for communicative teaching (Derwing, 2008; Thomson, in press), although it is appropriate to target those aspects of accent that impede intelligibility (Derwing & Munro, 2009b). In addition, accentedness is not included as an assessment criterion in the speaking scales of high-stakes tests used for academic purposes, such as the TOEFL or IELTS. In contrast, comprehensibility and fluency have been regarded as highly important for effective oral communication (Derwing & Munro, 2009a; Gatbonton & Segalowitz, 2005) and are included in the speaking rubrics of both L2 proficiency tests. However, L2 comprehensibility and fluency ratings were impervious to individual differences in rater cognitive abilities in this experiment.

In sum, in current thinking, accentedness is generally considered to be less important for L2 oral proficiency and for achieving effective communication than other

skills. Therefore, it is unlikely that the small difference detected in mean L2 accentedness ratings between music and non-music majors in this experiment (i.e., less than one scale point on the 9-point scale) would translate into differences in mean L2 oral proficiency ratings in real-world assessment contexts. Therefore, although the finding involving accentedness is of considerable research interest (e.g., with regard to which components of accent musical raters are more sensitive to), implications for assessment practice are limited. For instance, there is no indication, based on this study alone, that raters for high-stakes tests should be screened for musical ability or that a homogeneous group of raters should be sought based on their musical experience. Until future research suggests otherwise, language teachers and testers need not be overly concerned by this finding.

Since these particular rater cognitive variables appear to exert only a minimal influence at most on rater judgments of L2 speech, the next step in the research program pursued in this thesis was to examine other systematic sources of variance in the rating process. The two variables that are the focus of Study 2 are ESL teaching experience, which is a rater background characteristic, and rating scale length, which is a property of the rating scale. The rationale for examining these variables is that the 9-point Likert-type scales used in Study 1 have become conventionalized in L2 pronunciation research. Although Cronbach's alpha values using this scoring method are generally high (e.g., Derwing & Munro, 1997), it is unclear whether raters, and particularly novice raters, who lack experience formally assessing L2 speech, are reliably able to distinguish between nine distinct scale levels of L2 comprehensibility, accentedness, and fluency, particularly when the speech samples are relatively homogeneous in terms of oral proficiency and L1 background.

Even if experienced and novice raters do apply the scale levels consistently,

previous L2 assessment research has shown that raters assign the same scores for different reasons (Douglas, 1994; Turner & Upshur, 2002). Almost all L2 pronunciation studies involving raters report some measure of interrater reliability; however, few have probed raters' own interpretations of the constructs (e.g., Brown et al., 2005) or the strategies that they use to condense their complex impressions of the L2 performance samples into a single numerical rating (e.g., Lumley, 2005). In the absence of directive scale descriptors in the case of numerical rating scales, the onus is on each individual rater to come up with his/her own system for aligning scale levels with performance quality. The interpretative scope that raters are afforded using this procedure is considerable, and the extent to which raters converge in their understanding of the construct and its most salient features for making level distinctions can ultimately inform our understanding of the underlying L2 ability that we are attempting to measure.

It is perhaps fitting that the complex phenomenon of how raters distill L2 performances to arrive at scores be examined using multiple sources of evidence. Ultimately, it was the need to adequately address the research questions in this study that dictated the methods used. As in Study 1, Study 2 reports *t*-tests, correlations, and Cronbach's alpha coefficients, but extends these findings using additional data sources and analytic techniques (Rasch category probability plots, verbal protocols, and post-task interviews). Bringing together these different but complementary data provide a more complete picture of experienced and novice raters' approach to the listening and rating task from the raters' perspective and in the raters' own words.

Chapter 3 – Study 2

Rater Experience, Rating Scale Length, and Judgments of L2 Speech: Revisiting Research Conventions

Talia Isaacs, McGill University

Ron Thomson, Brock University

Isaacs, T., & Thomson, R. (under review). Rater experience, rating scale length, and judgments of L2 speech: Revisiting research conventions. *Language Assessment Quarterly*.

Rater Experience, Rating Scale Length, and Judgments of L2 Speech: Revisiting Research Conventions

Abstract

This mixed-methods study examines the effects of rating scale length and rater experience on listeners' judgments of second language (L2) speech. Twenty experienced and 20 novice raters, who were randomly assigned to either 5-point or 9-point rating scale conditions, judged speech samples of 38 newcomers to Canada on numerical rating scales for comprehensibility, accentedness, and fluency. Results yielded high interrater reliability coefficients and no group differences for rating scale length or rater experience. However, Rasch category probability plots revealed that raters had difficulty differentiating between scale steps, particularly in mid-scale range. Moreover, evidence from verbal protocols and post-task interviews suggested that experienced and novice raters adopted strategies to either draw on or offset their perceived experience with L2 speech in conducting their ratings. Implications for L2 speech research are discussed in light of current research conventions.

Keywords: rating scale, rater experience, oral assessment, pronunciation, mixed methods

Rater Experience, Rating Scale Length, and Judgments of L2 Speech: Revisiting Research Conventions

Introduction

Rating scales provide the framework within which human raters assign a score for a second language (L2) performance, which is taken to be an indicator of the L2 learners' ability on the trait or construct being measured. Such assessment schemes serve to constrain, structure, or filter raters' responses, often through rigid scale descriptors associated with a fixed number of scale bands. Lumley describes the tension between the "simplified orderliness of the rating scale," which necessarily underrepresents the complexity involved in L2 performance, and raters' unconstrained reactions to the performance, which may be disordered and complex (2005, p. 248). The challenge for raters is to reconcile their possibly idiosyncratic, intuitive, or non-linear impressions of an L2 performance with rating scale specifications, including deciding which fixed category of the scale a "grey area" performance fits into.

In L2 speech research, qualitative constructs of comprehensibility, accentedness, and fluency are most often measured on 9-point numerical rating scales (e.g., Derwing, Munro, & Wiebe, 1998; Kennedy & Trofimovich, 2008). Previous research on foreign accent scaling has argued for the use of at least 9-point scales to prevent a ceiling effect (Southwood & Flege, 1999), and this has been generalized to the use of 9-point comprehensibility and fluency scales (e.g., Derwing, Rossiter, Munro, & Thomson, 2004; Isaacs & Trofimovich, 2011). Generally, only scalar endpoints are defined in these Likert-type scales (e.g., *very difficult to understand* and *very easy to understand* for

comprehensibility). In the absence of descriptors in mid-scale range, raters may not have clear criteria for differentiating between scale steps.[1]

Despite this scope for interpretation, ratings derived using 9-point scales have consistently yielded high interrater reliability in empirical studies (generally Cronbach's alpha coefficients above .90), even for phonetically unsophisticated, untrained native speaking (NS) raters (e.g., Munro & Derwing, 1999). The scalar judgments in these studies are usually based on 20-30 s of recorded speech, since raters have been shown to reliably assess performances of this length. Further, Munro, Derwing, and Morton (2006) reported considerable agreement among raters from different first language (L1) backgrounds on the relative performance of L2 English speakers from diverse L1 backgrounds. This implies that impressionistic ratings assigned using this scoring method are somewhat generalizable across speakers and raters regardless of their L1 background.

Even when raters assign the same score to a speech sample, however, their rationale for doing so may be different. That is, while two sets of global pronunciation ratings may be quantitatively the same, this does not preclude qualitative differences in the raters' approach to the decision-making task or interpretation of the construct (Douglas, 1994; Upshur & Turner, 1999). Although there is a growing use of introspective techniques in L2 assessment research to examine how systematic sources of variability shed light on different dimensions of the construct (e.g., Barkaoui, 2007; Brown, Iwashita, & McNamara, 2005), the use of qualitative methods to probe rater processes as they listen to and score L2 pronunciation has been limited. Harding (2008) used focus groups and interviews to examine L2 learners' perceptions of different L1 English accents. The interest in the present study, however, is the reverse scenario—native English listeners' perceptions of L2 speech. Zielinski (2008) corroborated

observational data of NSs transcribing L2 speech with other data sources to reveal instances of unintelligibility. However, her sample size included only three listeners. Similarly, Rossiter (2009) examined rater motivations for scalar judgments of L2 fluency by asking raters to provide written notes prior to assigning ratings. However, her sample size was quite small (six in the case of expert NS raters) and the comments were relatively unelaborated. The goal of the present study, therefore, is to extend previous research by examining the effects of rating scale length, a method facet, and rater experience, a speaker characteristic, on rater decision-making in holistic assessments of L2 pronunciation.

Rating Scale Length

Ratings scales are commonly used in psychological research to operationalize constructs that cannot be directly observed and measured. The issue of the optimum rating scale length has been a research focus over the past several decades in psychological studies that have examined personal attitudes or preferences. Bendig (1953) found that reliability was stable in 3- to 9-point scales but was compromised when 11-point scales were used. Matell and Jacoby (1971) reported stable reliability in scales with 2 to 19 levels, concluding that increasing measurement precision by adding additional scale levels does not result in greater reliability. Thus, coarser scales (with fewer levels) were no less reliable than finer scales. Similarly, McKelvie (1978) found no psychometric advantage to scales with 10 or more levels, concluding that 5- or 6- point scales should be used. Despite the use of complex statistical techniques, the issue of the ideal number of rating scale categories remains unresolved (Preston & Colman, 2000). Overall, these findings can be roughly summarized by Miller's dictum, "the magical

number seven, plus or minus two: Some limits on our capacity for processing information" (1956, p. 81).

While including more scale levels should, in theory, allow finer-grained distinctions to be made between language performances, raters must be able to differentiate between all scale levels in order for measurement to be precise (Bachman, 1990). One goal of rating scale validation research, therefore, should be to investigate the number of scale levels that raters are reliably able to distinguish in their context of use. In the L2 assessment literature, there is some suggestion that raters have difficulty managing 9-point scales. The *Cambridge Assessment of Spoken English* scale, for instance, was reduced from nine to six band levels due to raters' "inability to differentiate effectively over all the scales" (Milanovic, Saville, Pollitt, & Cook, 1996, p. 19). Further, Hamp-Lyons and Henning (1991) suggest that "a nine point scale is longer than optimum for a writing test" due to the high cognitive load that is imposed when raters embark on the complex rating task (p. 364). They argue that the large number of scale levels makes attaining some degree of step separation difficult, noting that 6-point scales are the most commonly used in college writing tests. Finally, in reference to the precursor of the *International English Language Testing Service* speaking scale, Alderson (1991) described that the pronunciation content does not appear in all nine levels of the holistic scale, since it was thought that creating nine pronunciation levels might introduce artificial or unusable distinctions. This scale was subsequently redeveloped as an analytic scale, and the pronunciation criterion has recently been expanded from a 4-point to a 9-point scale (Develle, 2008). Part of the impetus for this revision was that the 4-point pronunciation scale was found to be too crude in its distinctions, and raters were not using the entire scale. To summarize, while language testers have long acknowledged that there

is no perfect scale (e.g., Underhill, 1987), the challenge is to develop a scale that is neither too fine-grained nor too coarse for a given assessment. Davidson (1991) suggests a role for scale step calibration using Rasch modeling to develop and refine rating scales.

In L2 accent research, rating scale length has been far from standardized (Piske, MacKay, & Flege, 2001). However, the 9-point numerical accent scale is becoming increasingly pervasive and has even been extended to comprehensibility and fluency judgments (e.g., Kennedy & Trofimovich, 2008; Munro et al., 2006; Rossiter, 2009). In an early study that employed the 9-point accentedness scale, Munro and Derwing (1994) reasoned that it was "better to overestimate the listeners' ability to resolve accentedness than to underestimate it" (p. 259). This echoes Flege and Fletcher's (1992) argument that using too fine a scale is preferable to using too restrictive a scale that fails to capture distinctions that listeners may make.[2] Southwood and Flege (1999) provided empirical backing for this view by comparing the adequacy of two accent scaling techniques. The first method derived accent ratings using a 7-point interval scale punctuated with the endpoints "least accented" and "most accented." The second method, direct magnitude estimation (DME), required raters to indicate the scope of the difference between a reference speech sample (baseline) and all other speech samples. A speaker judged to be twice as accented as the reference speaker, for example, would receive twice the score of the reference speaker; a speaker deemed half as accented would receive half the score. This method allows raters to construct their own ratio scale without being constricted by the endpoints imposed by an interval scale. The dispersion of DME scores led the researchers to conclude that 9- or 11-point interval scales are necessary to prevent a ceiling effect. The 7-point interval scale did not adequately reflect the magnitude of

differences that had been captured in the DME ratings, at least for a few raters who were sensitive to accent differences between early L2 learners and NSs.

The number of scale categories a rater is able to distinguish is not only constrained by his/her ability to detect differences between stimuli, but also by the discriminability inherent in the speech samples (Garner, 1960). The L2 speakers in Southwood and Flege's (1999) study were widely variable in their age of arrival (1.9–23.3 years) and length of residence in the target language country (14.6–44.3 years). Arguably, in most L2 speech research that does not deal specifically with age and accent, the range of such age-related variables tends to be much more restricted. In Derwing et al. (2004), for instance, the L2 speakers were adult immigrants who had resided in Canada for under 6 months. Similarly, the nonnative graduate students in Munro and Derwing (1999) had all learned English after puberty. Were Southwood and Flege's (1999) study to be replicated on either group of speakers, the detection of a ceiling effect on the 7-point scale might not be likely due to the presence of fewer native-like L2 speakers in the sample. Thus, it is unclear how well this ceiling effect would generalize to the samples more typically used in L2 speech research that does not examine age-related differences.

Rater Experience

In L2 assessment, experienced English as a Second Language (ESL) teachers have conventionally been called upon to make expert judgments of L2 performances or to validate rating scales (Brindley, 1991). Barnwell (1989), however, argues against the need to defer to teachers or expert assessors, since the domain of L2 oral proficiency lies outside the classroom. Because ESL teachers' and linguists' impressions are "atypical" compared to those of the interlocutors with whom L2 learners are likely to interact outside educational settings, Barnwell elaborates that naïve NSs should constitute another

"expert" audience that is consulted in rating scale validation. This opens up the definition of "expert raters," which has often been used interchangeably with the term "experienced raters" (e.g., Cumming, 1990). In L2 pronunciation and fluency studies, for example, expert raters have referred to a group of phoneticians and speech therapists (Cucchiarini, Strick, & Boves, 2002) and to ESL teachers with extensive teaching experience (Rossiter, 2009). Bongaerts, van Summeren, Planken, and Schils (1997) use the term "experienced judges" to include both phoneticians and English as a Foreign Language teachers, whereas Calloway's (1980) experienced raters were ESL teachers and teaching assistants. Finally, Kennedy and Trofimovich (2008) define experience as the degree of listeners' exposure to L2 speech, which in some studies is considered an indicator of rater familiarity. Clearly, the way that the expert or experienced rater and the nonexpert, inexperienced, naïve, novice, lay rater, or "person in the street" (Thompson, 1991, p. 177) are defined will impact the expertise/experience effect that is detected in a given study (Schoonen, Vergeer, & Eiting, 1997).

Definitional inconsistencies make cross-study comparisons difficult. Thompson (1991), for example, found that experienced raters with linguistic training and considerable contact with L2 learners were more reliable and "sympathetic listeners" than their "inexperienced" counterparts (p. 184). Conversely, Bongaerts et al. (1997) found no significant differences between experienced and inexperienced (non-linguistically trained) raters' accent ratings. More recently, Derwing et al. (2004) found that "untrained raters" (undergraduate students in an introductory TESL course) provided reliable judgments of L2 comprehensibility, accentedness, and fluency, and that their fluency ratings were strongly correlated with temporal measures of the L2 stimuli. For the purposes of the present study, experienced raters were defined as native English speaking

ESL teachers with postsecondary training, whereas novice raters were linguistically untrained NSs matched for education level.

If linguistically untrained listeners produce reliable ratings of L2 comprehensibility, accentedness, and fluency, then the often onerous task of recruiting nominally experienced listeners (however defined) might be unnecessary. Seeking ratings from "lay" raters is particularly attractive if their scores are statistically equivalent to those of experienced raters and they are representative of a segment of the people with whom the test-taker is likely to interact in real-world settings. However, even if there are no group differences in scoring outcomes, it may still be that experienced and novice raters approach the rating task in qualitatively different ways or focus on different rating criteria (Douglas, 1994).

The Present Study

Our interest in the numerical scales most widely used by convention in L2 speech research and in experienced and novice raters' scoring behavior led to the following research questions:

1a. Are there differences in the mean scores raters assign on measures of comprehensibility, accentedness, and fluency using 5-point or 9-point numerical rating scales?

b. Furthermore, is there evidence of rater preference for the 5-point or 9-point scale?

2a. Are there differences in the mean ratings assigned by experienced ESL teacher raters and novice raters on measures of comprehensibility, accentedness, and fluency?

b. Furthermore, do experienced and novice raters report using similar strategies to arrive at their ratings?

Method

Research Design

We employed a mixed methods triangulation design to address the research questions (Creswell & Plano-Clark, 2007). This involved converging different but complementary data (qualitative and quantitative) to shed light on the complexities involved in experienced and novice raters' judgments of L2 speech in the absence of detailed scoring rubrics.

Participants

The speakers were 38 adult newcomers to Canada (11 males, 27 females). Half were L1 Mandarin speakers and the other half were L1 speakers of a Slavic language (13 Russian, 3 Serbian, 2 Ukrainian, 1 Polish). They had arrived in Canada on average 15.6 months earlier (range: 2–42), had been assessed at the beginner levels of the *Canadian Language Benchmarks*, and were receiving ESL instruction through the Language Instruction for Newcomers to Canada program.

The raters were 40 adult L1 Canadian English speakers (11 males, 29 females) who either held or were pursuing postgraduate degrees. All reported having normal hearing and none had learned Chinese or Slavic languages. The raters were recruited to participate in experienced or novice rater groups. The experienced raters (6 males, 14 females) were 20 ESL teachers who had on average 9.68 years of teaching experience (range: 3–25) at an estimated 13.92 hr/week (range: 3–30). Despite being termed "experienced raters," they were variable in their teacher training. Thirteen had taken a course in phonology for teachers, 16 had taken an L2 assessment course, 10 had received over 2 hr of rater training, and two reported no pronunciation or assessment training. The novice raters were 20 graduate students (5 males, 15 females) from different academic

disciplines (e.g., geography, nursing, psychology) with no background in linguistics, assessment, or language teaching.

The experienced raters were significantly more familiar than novice raters with the spoken English of L1 Slavic, $t(28.78) = 3.50$, $p = .002$, and Mandarin speakers, $t(38) = 3.49$, $p = .001$, as assessed on 9-point familiarity scales (1 = *extremely unfamiliar*, 9 = *extremely familiar*). The mean familiarity rating for Slavic speakers was 4.35 for experienced ($SD = 2.43$) and 2.20 for novice raters ($SD = 1.28$). The mean familiarity rating for Mandarin speakers was 5.60 ($SD = 2.39$) for experienced and 3.15 ($SD = 2.03$) for novice raters. While rater groups did not differ in the percent of time they reported speaking or listening to languages other than English, experienced raters reported significantly more interaction with L2 speakers than novice raters, $t(38) = 3.02$, $p = .005$, some of which presumably took place in the ESL classroom itself. Experienced raters estimated interacting with L2 speakers 39% of the time ($SD = 16.83$) compared to 22.5% for novice raters ($SD = 17.73$).

Procedure

Speech Elicitation and Stimulus Preparation

L2 speech samples on six speaking tasks were elicited individually in 1 hour sessions. This paper will examine performance on one task, an eight-frame picture narrative that has been widely used in previous L2 speech research (e.g., Derwing, Munro, Thomson, & Rossiter, 2009; Trofimovich, Gatbonton, & Segalowitz, 2007). The picture sequence featured a man and a woman who bumped into each other and dropped the identical suitcases they were carrying, only to realize, upon reaching their respective destinations, that they had inadvertently exchanged suitcases.

After normalizing the speech samples, the first 20 s of each narrative, excluding initial dysfluencies, were burned onto CDs in three randomized orders. Following Derwing, Thomson, & Munro (2006) and Derwing et al. (1998), the speech sample of a male native English speaker on the same task was included approximately two thirds of the way through each randomization to ensure that the scores the raters assigned corresponded to the correct item number (speech sample). This strategy assumes that the NS will be rated at the high end of the scale. Once it was determined that the speech sample and rating item number corresponded, the NS' ratings were discarded from all subsequent analyses.

Experimental Conditions for Raters

Of the 20 experienced and 20 novice raters, half were randomly assigned to either the 5-point or 9-point scale conditions, with 10 raters per condition. The raters in the 5-point condition rated the speech samples on separate 5-point numerical scales for comprehensibility (1 = *very hard to understand*, 5 = *very easy to understand*), accentedness (1 = *heavily accented*, 5 = *not accented at all*), and fluency (1 = *very dysfluent*, 5 = *very fluent*). The raters in the 9-point condition assigned scalar judgments on the same three constructs using three separate 9-point numerical scales with the same descriptors at the scale anchors, with the difference that '9' (rather than '5') was the scalar endpoint designating the highest level of performance. That is, the only difference between the rating conditions was the number of scale levels.

We employed verbal protocols to shed light on experienced and novice raters' scoring decisions. The rationale was that raters' verbalizations might offer an indirect, albeit incomplete glimpse into their internal thought processes while they reflected on and evaluated the L2 speech (Ericsson & Simon, 1993). However, the artifice of having raters

think aloud while rating has the potential to obscure precisely the phenomenon it was being used to probe, namely, the nature of the rating process (Lumley, 2005). Therefore, it was necessary to build into the research a means of investigating the effect of the additional demand of having raters think aloud on their scoring outcomes. To this end, half of the experienced and novice raters assigned to both rating scale conditions were, in turn, randomly assigned to either a verbal protocol (think-aloud) condition (n = 5) or a rating only (no verbal commentary) condition (n = 5).

The Rating Sessions

The rating sessions, which were conducted individually, lasted, 1–2 hours. This included a training component and a break to mitigate rater fatigue. In order to ensure consistency in the presentation of the constructs, typical construct definitions based on previous research were provided. Comprehensibility was defined as "how easy the speaker is to understand." Accentedness denoted "how different the speaker sounds from a NS of North American English" (Munro & Derwing, 1999). Fluency, here defined temporally (Lennon, 1990), was described as "how smooth the speaker's oral delivery is based on pausing, hesitation markers, fillers (e.g., um, uh), or speech rate."

After familiarizing themselves with the speaking prompt and receiving oral reinforcement on written instructions, the raters judged five practice speech samples for comprehensibility, accentedness, and fluency on three separate 5- or 9-point numerical rating scales, depending on the condition.[3] In addition, raters in the verbal protocol condition practiced thinking aloud. They subsequently received feedback on each rating by comparison with mean scores assigned by an independent group of raters from a previous study (Derwing et al., 2009). This brief calibration was designed to familiarize raters with the overall speaking ability to expect and to give them a rough idea of

previous raters' assessments. A NS was included in the training session to establish the upper bounds of the scales.

The verbal protocols took, on average, 41.1 min for the experienced raters and 34.4 min for the novice raters to complete (27–61 min). At the end of each speech sample, the recording was paused and the rater scored the speech and verbalized his/her thoughts on the speech and rating process. In the event of a sustained silence, the researcher prompted the rater to continue verbalizing (Gass & Mackey, 2000). However, the rater was ultimately the arbitrator of how much to say, with the researcher typically picking up on verbal and non-verbal cues that the rater was ready to move on. Conversely, listeners in the rating only condition scored the speech without providing verbal commentary during a 7 s interstimulus interval. The ratings in this timed condition took approximately 18 min to complete.

Following this, all raters participated in a second listening while looking at their original ratings and, when the recording was paused, articulated thoughts about the rating process or their impressions of the speech. This served as a check on the consistency of the verbal reports of raters in the protocol condition (Ericsson & Simon, 1993) and a means of probing the impressions of raters in the rating only condition, who had hitherto remained silent. Finally, all raters were interviewed about their impressions of the task, interpretations of the constructs, scoring behaviour, rating strategies, and influences on their assessments. They were specifically asked to comment on their scale use and criteria for distinguishing between low and high ability speakers.

Data Analysis

Classical statistics were conducted using *SPSS* 17. In preparation for analysis, the comprehensibility, accentedness, and fluency scales were normalized by scaling the 9-

point scale down to a 5-point scale using the equivalencies shown in Table 1. Because a wider-ranging scale typically yields larger variance than a smaller-ranging scale, changing the metric of the 5-point scale so it would encapsulate the same numerical range as the 9-point scale and adjusting the ratings accordingly made the variance of the two sets of ratings more amenable to parametric comparison.

Table 1. *Normalization of 9-Point Scale to 5-Point Scale*

	Scale points								
Original 9-point scale	1	2	3	4	5	6	7	8	9
Normalized 9-point scale	1	1.5	2	2.5	3	3.5	4	4.5	5

In order to gain insight into how raters applied the 5- and 9-point scales, Rasch analyses were conducted on the original (unnormalized) rating data using *FACETS* 3.60 (Linacre, 2005). This program extrapolates estimates of speaker ability level and rater severity based on the scores that individual raters assign to each speech sample. It then maps these facets of "speaker" and "rater" on a logit (arithmetic) scale, which provides a common yardstick for examining these measurement components. In this study, separate analyses were conducted for comprehensibility, accentedness and fluency for both rating scale conditions (3 Perceptual Measures x 2 Rating Conditions), since these were rated on separate scales with different metrics. In total, six analyses were conducted with 760

ratings calibrated for each analysis (38 Speakers x 20 Raters). While most Rasch modeling studies in L2 assessment have focused on rater severity, item fit, or bias detection (e.g., Johnson & Lim, 2009), this paper will limit its discussion to scale response category plots. This tool has been used to inform the extent to which raters distinguish between scale points on multistep interval scales (e.g., Milanovic et al., 1996).

The think-aloud and interview data were analyzed using *ATLAS.ti* 5.0.66, a qualitative software package that facilitates categorizing data and making associative links between multiple data sources. In an iterative process, the verbal protocol data were segmented, coded, and mapped onto the interview data to generate emergent themes (Green, 1998). Segmented episodes included rater comments about their role as raters, scoring decisions, and observations of the speech, rating task, or think aloud. Rater comments from the verbal protocol and interview data pertaining to the research questions will be reported here.

Results

Internal Consistency, Correlations, and Comparison of Means of Experimental Subgroups

In our first analysis, we computed Cronbach's alpha to examine the internal consistency (homogeneity) of the comprehensibility, accentedness, and fluency ratings. Table 2 shows that Cronbach's alpha values exceeded .90 for all experimental conditions. Judgments obtained using the 9-point scales were slightly more consistent than those obtained using the 5-point scales, and the coefficients for experienced (ESL teacher) raters were only marginally higher than for novice raters. Increases in Cronbach's alpha when all 40 raters are pooled evidences the sensitivity of this measure to sample size.

Table 2. *Cronbach's Alpha for Comprehensibility, Accentedness, and Fluency for Each Experimental Condition and Pooled Across Conditions*

		Comprehensibility	Accentedness	Fluency
Rating scale length	5-point[a]	.92	.92	.94
	9-point[a]	.95	.94	.95
Rater experience	Experienced[a]	.95	.94	.95
	Novice[a]	.92	.92	.94
Verbal protocol	Protocol[a]	.93	.92	.95
	Rating only[a]	.93	.93	.95
	Pooled[b]	**.97**	**.97**	**.97**

[a]$n = 20$; [b]$n = 40$

Next, Pearson correlation coefficients were computed to examine the shared variance between the 5- versus 9-point scales, experienced versus novice raters, and protocol versus rating only conditions. Table 3 shows that raters reached considerable consensus about speakers' ability regardless of experimental condition, with shared variance of at least 90% in all cases.

Table 3. *Pearson's Correlation Coefficients of Ratings of Experimental Subgroups*

	Experienced vs. novice	5- vs. 9-point scale	Protocol vs. rating only
Comprehensibility	.92**	.90**	.95**
Accent	.92**	.91**	.94**
Fluency	.96**	.94**	.96**

****** $p < .001$, two tailed

Having established high overall agreement for each experimental condition on all dependent variables, we averaged the scores for each experimental subgroup over the 38 items (speech samples). We then conducted three-independent *t*-tests to examine group differences between (a) the 5- and 9-point scales; (b) experienced and novice raters; and (c) protocol and rating only conditions. A two-tailed Bonferroni correction was applied to adjust the significance level for multiple comparisons. Results revealed no significant differences on any of the dependent variable measures for rating scale length, rater experience, or protocol condition, $t(74) = -2.1–.80$, $p > .017$. That is, the scores assigned by the subgroups were not quantitatively different.[4] In the remainder of the paper, evidence from additional data sources will be used to extend the null result for rating scale length and rater experience.

Rating Scale Use and Preference

As a follow up to the nonsignificant *t*-test result for scale length, we calculated each rater's mean skewness score for comprehensibility, accentedness, and fluency. The purpose was to examine whether differences exist in the centeredness of the distribution for ratings obtained using the 5- and 9-point scales. *T*-tests revealed no skewness differences between the rating scale conditions. Notably, all distributions were slightly positively skewed. The skewness coefficients for fluency, which were the largest of the three measures, were .36 for both rating scale conditions, which is within the range of distribution normality (Huck, 2004). Figures 1–3 show that when ratings are pooled over experience and protocol conditions, the score distributions for the 5- and 9-point conditions mirror each other. This again suggests no differences in raters' overall scoring patterns based on rating scale condition.

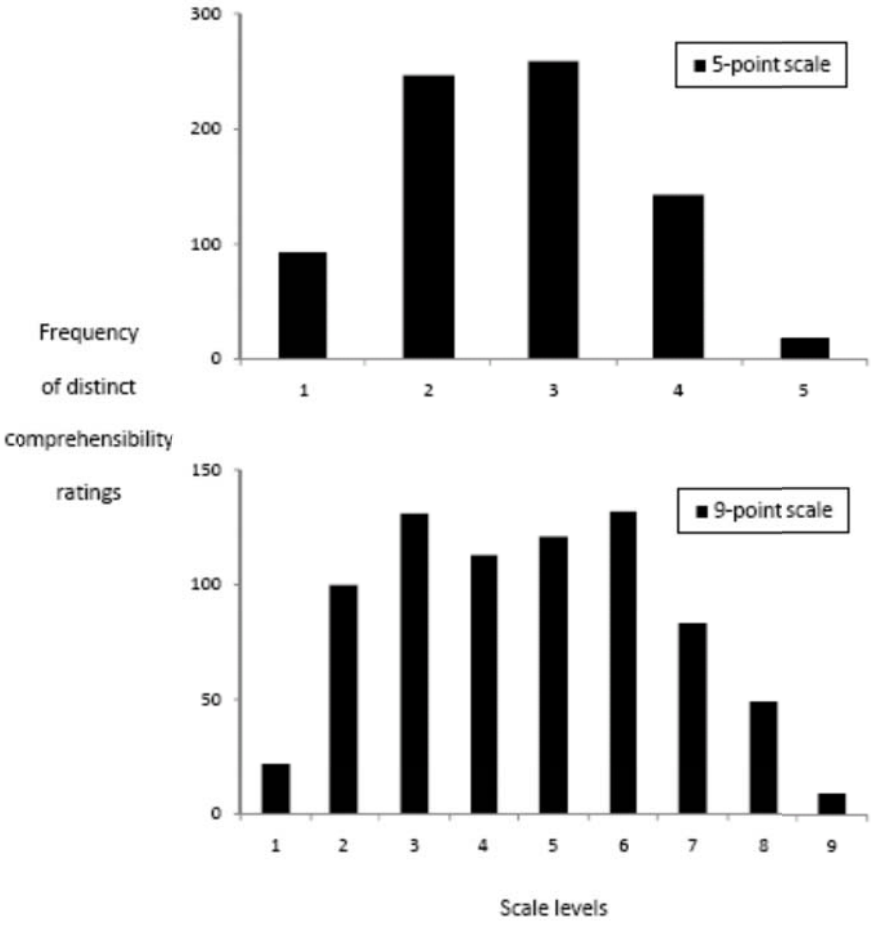

Figure 1. Score distributions for 5- and 9-point comprehensibility sales

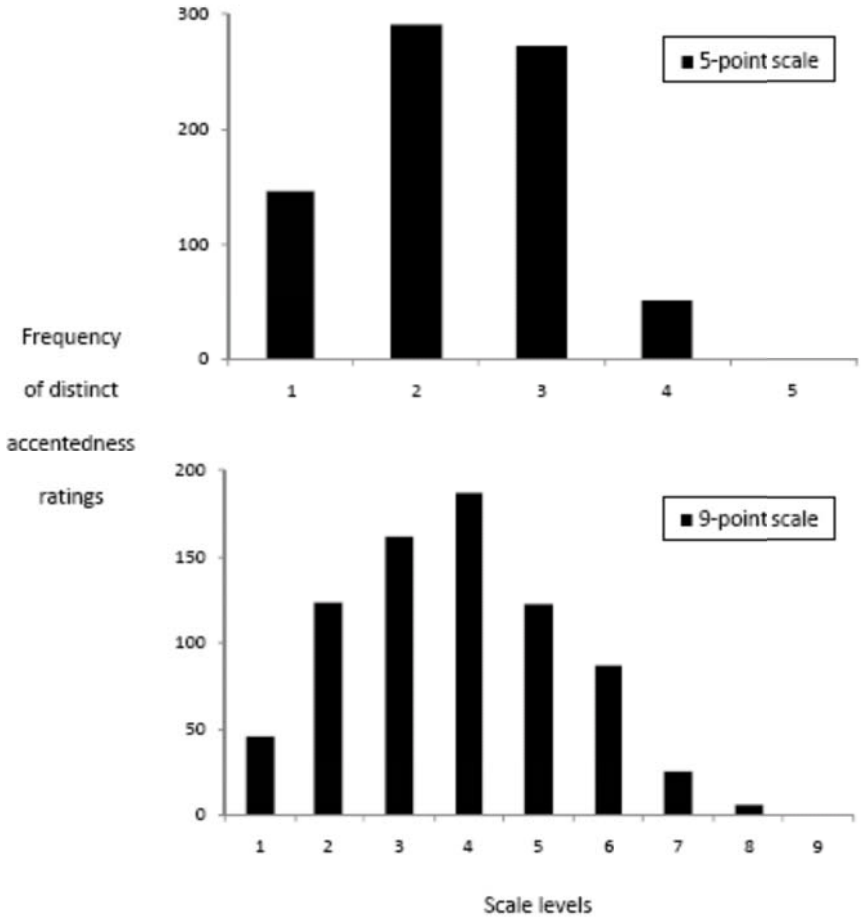

Figure 2. Score distributions for 5- and 9-point accentedness scales

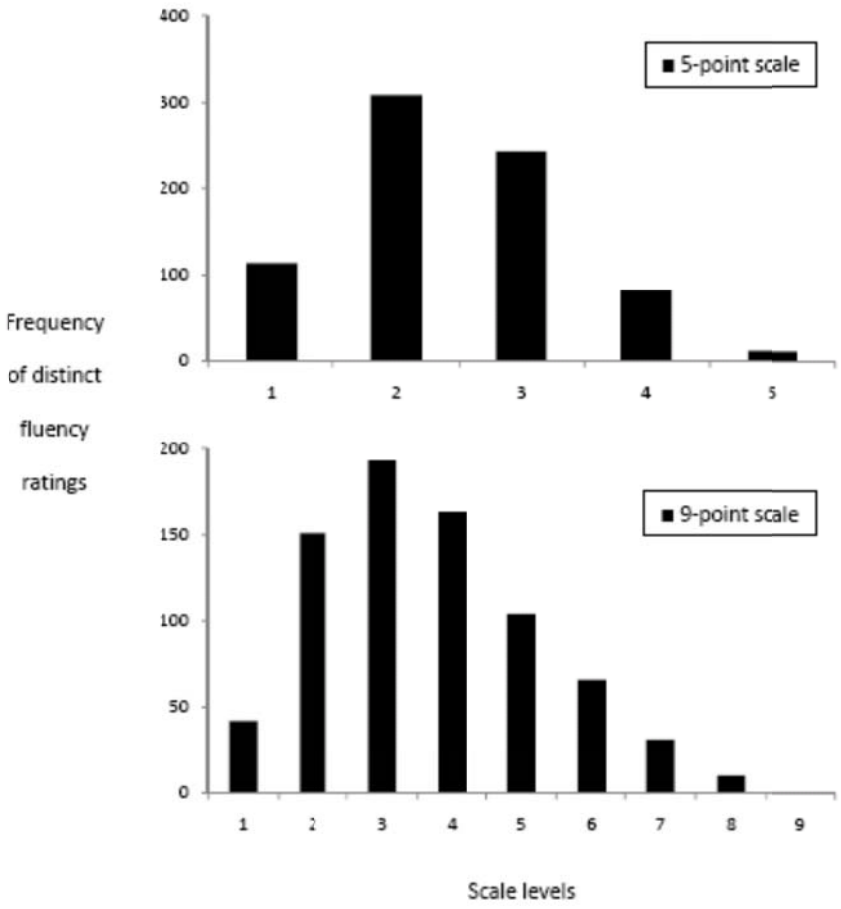

Figure 3. Score distributions for 5- and 9-point fluency scales

Notably, no L2 speakers received a score of '5' on the 5-point accentedness scale or '9' on the 9-point accentedness or fluency scales. Several raters echoed the observation by Nov.R8.9pt (Novice Rater 8, 9-point scale condition) that "there's certainly a big gap between the native English speakers and almost everyone in this study. If a NS is 9, these people certainly weren't even close to being a 5." A majority of raters revealed reserving the upper bounds of the accentedness scale for speakers who approached the NS standard. This was in accordance with the definition of accentedness as the deviation of the accent from the standard language norm coupled with the use of NSs to establish the upper

bounds of the scale during rater training. In fact, 11 raters in the 5-point condition only worked with a 4-point accentedness scale, and the remaining nine raters "were just using from 3 to 1, because 4 to 5 were so high," as Exp.R21.5pt (Experienced Rater 21, 5-point scale condition) stated.

Several raters expressed discomfort rating accentedness. Nov.R38.9pt articulated the prevalent view that "accentedness is amorphous in the sense that as long as I can understand the word, I don't mind if it's said with a non-North American accent." In some cases, raters' beliefs about the unimportance of accentedness relative to comprehensibility even led to indiscriminate scoring. Exp.R3.5pt, for example, related, "I put it [accentedness] down in the middle as something that doesn't really matter. So my default was a 3 for accent. Yes he had an accent but it didn't matter." In fact, R3 assigned scores of 3 for over 70% of her accentedness ratings but for only 34% of her comprehensibility and fluency ratings. By not distinguishing between speakers on accentedness in a study where comprehensibility, accentedness, and fluency were implicitly assigned the same importance, R3 was arguably successful at "putting accent in its place" (Derwing & Munro, 2009, p. 1).

When asked if they had any difficulties using the rating scales overall, some raters expressed that the 5-point scale was too constraining in that they tended to want to assign half points in the lower scale range. Exp.R36.5pt described that, at the onset of the speech sample, he would circle a number based on his initial judgment but then would write an arrow as the speech progressed to signal whether the actual score was slightly higher or lower than he had indicated. He noted, "you have only five categories, but it could've been a thousand categories actually." Although R36 was alone in adopting this strategy, some raters clearly saw more subtleties in the performances than a 5-point scale could

accommodate. In contrast, proponents of the 5-point such as Nov.R24.5pt noted that shorter scales are more "meaningful" and that "fewer points is easier to rate because you just have less nuance."

Opinion on the 9-point scale was also mixed. However, a majority of respondents reportedly found the 9-point scale difficult to manage, particularly in mid-scale range. Exp.R39.9pt stated, "I think it's hard for people to set real discrete categories. I know I had trouble with the middle [of the scale]. I think sometimes it's a coin toss." Exp.R5.9pt, whose verbal protocol was the longest, frequently appeared conflicted about score assignment. An example of a scoring decision episode from his verbal protocol is, "I feel like I'm rating more objectively now. She's maybe at a 4, 5... 4. That's pretty hard to understand. That's a pretty strong accent, 5. Maybe 4. But I'll leave it at 5." In the follow-up interview, R5 underscored his difficulty arriving at a scoring decision, stating, "I'm wondering how accurate this would be, because I just feel so uncertain while I'm doing it. And there are so many times that I'm like I'll just go with a 6, because I can't figure it out. Is it 7 or 6? I don't know man. I'm just going to go with the 6… It's annoying because you don't get the resolution, so it's very frustrating, because you're like I just don't know. Is it a 6 or a 7? So you get frustrated." Ironically, when asked about using the scale, R5 replied, "I like that there are nine [categories], because you can really play with the 7, 8, 9." Thus, despite feeling uncertain about his scale step choice, R5 liked the scope offered by the 9-point scale. It is possible that his scoring decisions would have been as laboured had he been assigned to the 5-point condition. However, the interviews revealed that no raters in the 9-point condition came up with a system to differentiate all adjacent scale levels. In fact, although raters were instructed to use the entire scale range, none in the 9-point condition did so, whereas 12 out of 20 raters in the 5-point condition made use

of all five scale points for either comprehensibility or fluency. In fact, two raters in the 9-point condition used only a 5-point scale and eight used just a 6-category range for comprehensibility or fluency. It is likely that this restricted scale use was, in part, an artefact of the low L2 ability of the speakers in this study.

In order to gain further insight into raters' application of the scales, Rasch probability plots were examined. Figures 4 and 5 show the likelihood of a given rating scale category being assigned across the comprehensibility ability continuum on the 5- and 9-point scales respectively. Response category probability curves for comprehensibility are shown, since both rating scale groups used the full range of the comprehensibility scale. Each step calibration increases monotonically with the scale step number. That is, as speaker ability increases, so does the probability of the speaker receiving a higher score (Linacre, 1999). Outfit mean-square values for the scale categories are close to the expected value of 1.0 for both scales (range: 0.9–1.1), which signifies that all scale categories meet Rasch model specifications.

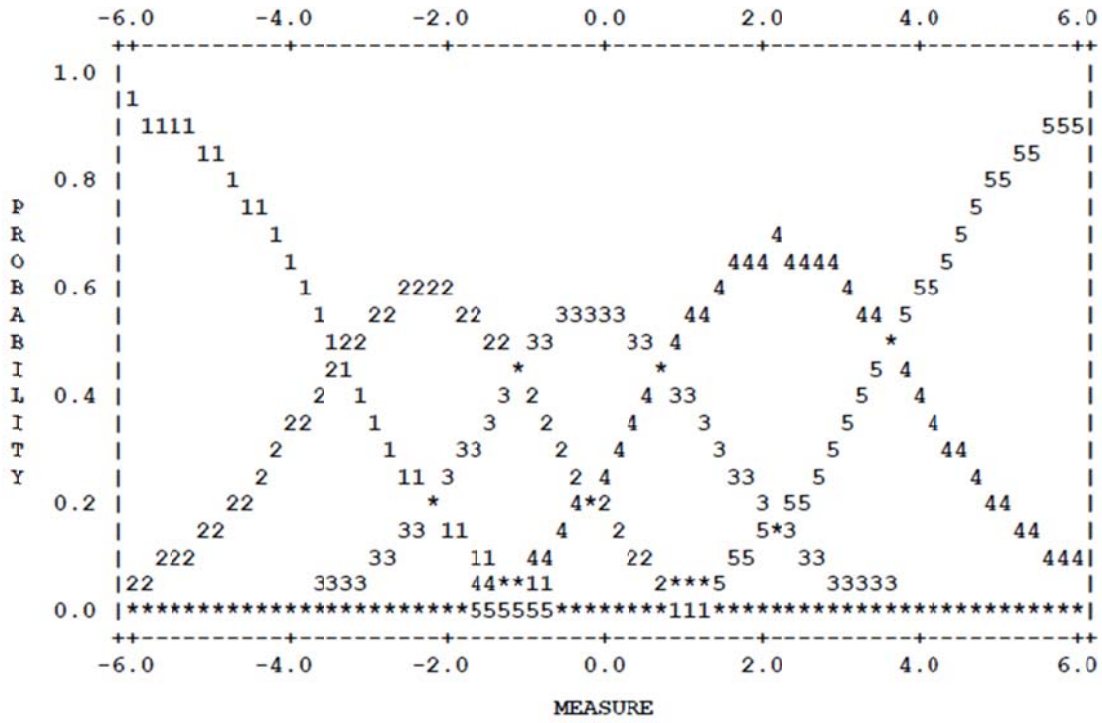

Figure 4. Response category probability curves for comprehensibility, 5-point scale

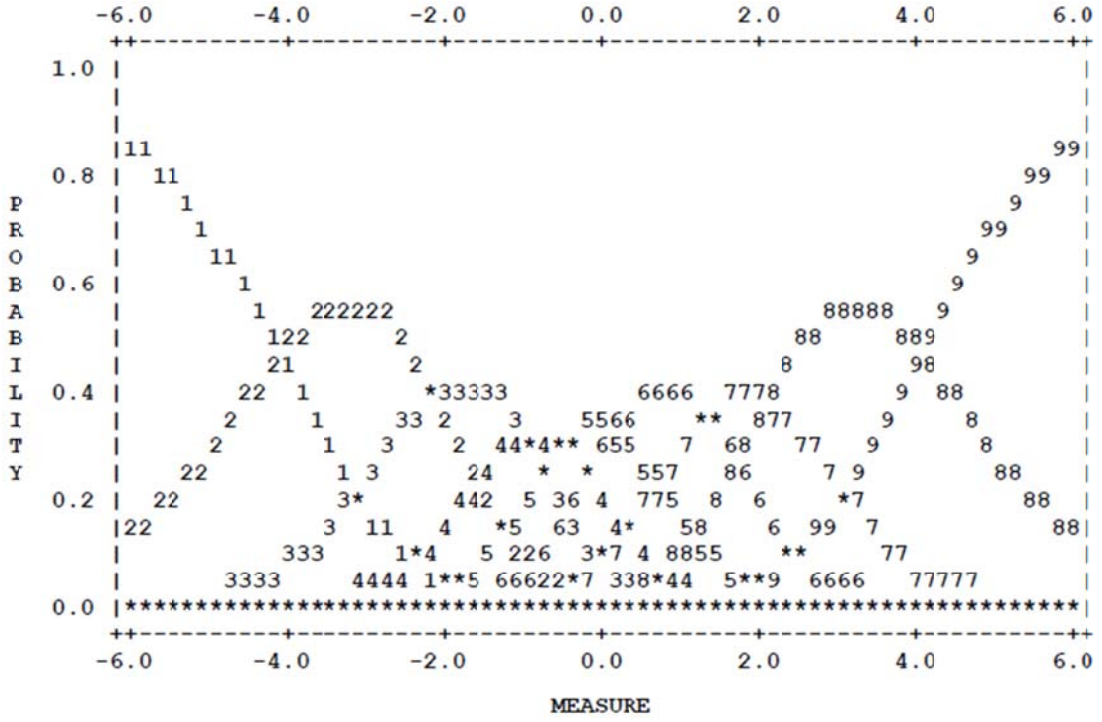

Figure 5. Response category probability curves for comprehensibility, 9-point scale

In probability plots for polytomous data, the categories at the scalar extremes always approach a probability of 1. In an ideal calibration, the probability curve for each step between the scalar extremes should have a high peak and cover a distinct area on the ability scale with minimal overlap (Davidson, 1991). Clear separation between adjacent levels should be visually apparent. As Linacre (1999) describes, "most scale developers intend this [probability curves] to look like a series of hills" (p. 138). If peaks appear at approximately the same level, then the maximum probability of receiving each score is similar for the various levels.

Evidently, the data in this study fall short of ideal. Figure 4 shows some separation between levels of the 5-point scale, in that one scale category is clearly the most probable at all points on the ability continuum. The probabilities associated with scale steps 2 to 4 range from 55% to 62%. However, the peaks, while easily discernable, are not very steep. Further, there is considerable overlap between the probability curves by more than just the neighboring category. A considerable portion of what should have been in the "jurisdiction" of scale level 3, for example, is encroached upon by levels 2 and 4. Notably, scale category 4 is higher peaked and more expansive than categories 2 and 3. However, this spread indicates less measurement precision and, therefore, a greater chance of inappropriate scoring (Milanovic et al., 1996). This may have been due to fewer instances of response category 4 in the dataset relative to categories 2 and 3 (see Figure 1).

Overall, it appears that the 5-point scale was applied with some clarity, although level distinctions were not always discernible to the raters. Limited rater training and the lack of external guidance on the quality of speech characterizing each scale level likely contributed to this. Indeed, as Exp.R37.5pt observed, the rating task in this study

mandated that raters create their own scales. It is possible that they each developed a distinct system of selecting scale categories in response to the speech, although how explicitly they created and adhered to their own distinctions is unclear. The interpretative scope raters were allowed in this study is reflected in the overlapping categories of the probability curves.

Figure 5 shows that the scale categories of the 9-point scale are less discernable than those applied using the 5-point scale. While scale levels 2 and 8 are nicely peaked, the "hilltops" of levels 3 through 7 are not easily distinguished, with the probability of occurrence for those categories ranging between 32% and 42%. In fact, 76% of the total comprehensibility scores assigned were between scale points 3 and 7, yet there is less than a 50% chance that a speaker in this ability range will be assigned the most probable score. Thus, it is possible that a speaker meriting a score of 2 based on the model calibrations will actually receive a score of 5. Scale levels 4 and 5 are overshadowed by heavily overlapping adjacent levels, which, again, leaves little assurance that a speaker at the corresponding ability levels will be accurately assessed. Clearly, the categories in the midsection of the scale are muddled, which lends credence to Nov.R31.9pt's view that "it [the scale] was a bit too long to gather better precision" and "it seemed like there's an extra point in there." One approach might be to collapse categories 3 and 4 or 5 and 6. Milanovic et al. (1996) suggest that raters may have more difficulty judging mediocre performances than extreme ones, and this seems to have been the case for the raters in this study.

Experienced and Novice Raters' Scoring Tendencies

While the *t*-tests revealed no mean differences between experienced and novice raters, some group differences were apparent in the qualitative data. One difference was

that experienced raters' verbal reports and interviews were about 5 min longer than those of novice raters on average, although no significant length differences were detected. Another difference was that some experienced raters expressed that ESL teachers are more adept at understanding L2 speech than "the typical NS" or "average person whose only qualification is they speak English" (Exp.R37.5pt). For instance, Exp.R3.5pt revealed needing to put her "best teacher ears on to understand" the speech of a low rated speaker. Further, Exp.R5.9pt, in reflecting on the rating process, divulged, "initially I was only in my shoes [as an L2 teacher], but by the end I was like this person is hard to understand, I was in the average person's shoes. So I was oscillating between that." Exp.R14.5pt expressed attempting to "prevent myself from using my prior knowledge to overinterpret what they're saying and fill in the gaps" as a "mental manoeuvre that required some concentration." This involved trying to disregard having seen the picture prompt when scoring, a strategy also adopted by some novice raters. In addition, R14 tried to consider only the "general ability for the [non-ESL teacher] interlocutor" to process the speech when assigning scores and described this attempt to trump his own experience as "an irony in my filling out of the scale." This strategy of trying to make his ratings resemble lay ratings was uniquely adopted by experienced raters. Conversely, some novice raters like Nov.R10.5pt conceded, "I don't have any experience describing this kind of thing [L2 speech], so I'm like uhh." Nov.R19.9pt challenged her rating qualifications due to her perception of her own less than articulate verbal discourse, commenting, "like who am I to evaluate someone?" This may have been a by-product of self-consciousness about thinking aloud, a pervasive sentiment in transcripts of both experienced and novice raters. The fact that several experienced raters constructed their identities based on their experience listening to, evaluating, or deliberating about L2

speech and occasionally tried to discount this experience, whereas several novice raters acknowledged their lack of experience, even if they had frequent interactions with L2 speakers, validates the a priori distinction made between experienced and novice raters in this study.

Experienced raters frequently referred to the voices in the speech samples as "students," which implies that they extrapolated the research context to their classroom experience. Some related anecdotes about strategies they teach their students to overcome the challenges they felt were shared by the speakers in the recordings. Others judged the rating task or quality of the scale using the benchmark of its usefulness for the classroom. Novice raters, on the other hand, who lacked ESL students as a point of reference, instead referred to the speech of L2 contacts from their inner circle (e.g., Austrian relatives, Chinese co-worker, Ukrainian roommate) or individuals from popular culture (e.g., Arnold Schwarzenegger, Borat) in conducting their ratings. Thus, it appears that experienced and novice raters used different reference points in evaluating L2 speech yet arrived at virtually identical scoring decisions.

The experienced raters varied in their access to technical vocabulary to describe L2 pronunciation specifically. Exp.R39.9pt, for example, was familiar with the term "vowel epenthesis." In contrast, Exp.R13.5pt, who reported no phonological training, referred to the "pronunciation of certain *letters*," whereas an individual with such training likely would have referred to the pronunciation of phonemes. The novice raters, who more uniformly lacked access to technical vocabulary and ESL jargon, tended to be more creative in describing the speech. For example, "flashcard speaking" referred to a lack of linking, "dead air" denoted silent pauses, and "noise" designated filled pauses for three novice raters. Nov.R15.5pt, who was the most vivid in her descriptions, recalled her

impressions of one voice as sounding like "you put a clock over it [the speech] to dampen the English pronunciation," whereas another speaker's "slow" and "heavy" speech "felt like moving in water." In sum, while differences between experienced and novice raters were masked in the quantitative analysis, the qualitative data accentuated group differences due to L2 teaching experience. Novice raters tended to compensate for their lack of technical vocabulary through reference to the speech of familiar L2 speakers and, in some cases, through creative descriptions.

Discussion

This study examined the effects of rating scale length and rater experience on judgments of L2 comprehensibility, accentedness, and fluency. Qualitative data were used to extend quantitative findings. There were no differences in mean scores obtained using the 5- versus 9-point scales, and high Cronbach's alpha coefficients were yielded for both conditions. However, Rasch probability plots revealed considerable overlap between scale categories for the 5-point and particularly the 9-point scale, which suggests that raters did not clearly differentiate between scale categories. Stemler and Tsai (2008) note that Cronbach's alpha is less stringent than other interrater reliability indices, since high coefficients are yielded as long as each rater "is consistent within his or her own definition of the rating scale" (p. 42). When group ratings are pooled in this procedure, the idiosyncrasies associated with an individual rater's scale use get averaged out. In this study and previous L2 speech research that employs 9-point scales, the derived scores are not being used to make high-stakes decisions. Rather, the ratings are solely being used to compare participants' performance for research purposes. In this context, raters' exact application of the scale is unimportant, provided that there is agreement on the overall ranking of L2 speakers. Clearly, it would be more important to achieve conformity on the

meaning and use of scale levels in operational assessments, particularly when rating outcomes have the potential to influence high-stakes decision making. Nonetheless, raters conducting impressionistic ratings in low-stakes research contexts may benefit from a more concrete understanding of the "no man's land" of ability levels between scalar extremes (Exp.R14.5pt).

While the 5-point scales were reportedly too constraining for some raters, particularly at the low end of the scale, it appears that the 9-point scale was difficult for raters to manage, in that there were "so many numbers" that raters were unable to meaningfully differentiate between or use all scale levels (Nov.R19.9pt). In contrast to Southwood and Flege's accent scaling study (1999), there was no evidence of a ceiling effect. Clearly, the distribution of scores is dependent on the ability levels manifested in the speech samples. Due to the low L2 proficiency of the speakers in this study, a floor effect would have been more probable than a ceiling effect, even though some raters admitted reluctance to assign scores of '1' because "there's a certain moral failing attached to the low end of the scale" (Exp.R14.5pt). To summarize, one implication of the findings is that rating scale use is clearly sample dependent. It may be that Southwood and Flege's arguments for using a 9- or 11-point interval scale for measuring L2 accent should not be generalized to studies that draw on more homogenous samples of L2 speakers (as was the case in the present study). Indeed, without rigorous training and calibration of raters at the start of rating sessions, it is unlikely that any speech sample from this study would receive an equivalent mean rating if it were presented within a more heterogeneous set of speech samples. For example, a speaker in the present study might be assigned a rating of '6' on the 9-point accentedness scale but a '3' in a study where more of the speakers had less obviously nonnative accents. This suggests that, in

the absence of rigorous rater training, absolute ratings are only interpretable within a specific study, not across studies.

Some raters were uncomfortable with the NS standard that loomed over the study, an issue that arose in accentedness judgments in particular. ExpR14.5pt expressed, "I think you need a scale… that allows successful nonnative speakers to be at the top end of the scale. We need to make sure that the scale reflects our judgments about success and not judgments about who their parents are or their first language. If you detect a first language, that should not put them lower on the scale." While accentedness is typically defined in reference to a NS standard, using NSs to establish the upper bounds of the scale during rater training likely reinforced this unrealistic standard and may have discouraged more comprehensive scale use. Rating scales in operational L2 speech research should perhaps more explicitly and substantively define the ability range that they are measuring and that is appropriate for the population being assessed in light of the language domain that is being generalized to. It is unlikely that the newly-arrived immigrants in this study, for example, needed to sound like NSs in order to integrate into society or successfully communicate in the workplace (unless, perhaps, they were aspiring intelligence agents). The implication is that even in research settings, raters should be made aware of the assessment context so that their judgments are more specific to the particular sample of L2 speakers being evaluated. This is an important consideration for future L2 speech research that uses a similar methodology.

While there were no significant differences in experienced and novice raters' scoring outcomes and interrater reliability coefficients for both groups were high, the qualitative data revealed differences in raters' approach to the task. Several experienced raters believed that their experience with L2 learners might affect their comprehension

and appraisal of the speech relative to non-ESL teachers. Some even attempted to put themselves in the place of a non-ESL trained interlocutor who was unfamiliar with the picture prompt when assigning scores. Conversely, novice raters, who were less schooled at describing L2 speech, tended to compensate for their dearth of ESL jargon or technical vocabulary by using more creative descriptions of the speech patterns or alluding to the L2 speech of personal acquaintances or celebrities. To summarize, both experienced and novice raters adopted strategies to either draw on or offset their perceived experience with L2 speech in conducting their ratings.

Despite qualitative differences between experienced and novice rater groups, the absence of significant quantitative differences in their scoring suggests that choosing novice raters over experienced raters, at least as defined in this study, might be justified when mean ratings are solely being used for research purposes. Such a scenario might arise if it is easier to recruit novice raters than experienced raters. It also appears that an interrater reliability measure that pools ratings might yield similar coefficients for both groups. This is perhaps unsurprising, given the minimal rater training in this study and that the descriptors at the scalar extremes, although vague, used nontechnical terms interpretable to both rater groups. In sum, when the interest is to examine speaker performance based on mean scores and not on qualitative insights into rater decision-making, there seems to be little advantage to obtaining experienced raters' judgments over those of novice raters. However, in higher-stakes assessments, garnering evidence that raters are attending to the construct of interest is likely to be paramount for validity reasons (see McNamara & Roever, 2006). In the present study, it is unclear if novice raters lacked the vocabulary of experienced raters and, therefore, had difficulty verbalizing their perceptions of the speech, or rather if experienced and novice raters were

attending to qualitatively different dimensions of the speech overall. What is clear is that in both high-stakes settings and research contexts, raters are likely to benefit from clearer operationalization of the construct, including guidance about what numerical scale points "mean" in terms of performance quality (Fulcher, 1996). For this reason and in order to refine our understanding of the construct, it is important to examine in greater depth what raters attend to when scoring. Brown et al. (2005) have spearheaded work on this in the assessment of speaking, although little L2 assessment research has focused on pronunciation specifically.

Concluding Remarks

As Lumley (2005) notes, "the search for the perfect scale (the 'holy scale') is futile" (p. 301), since scales are imperfect representations of reality and underrepresent the complexities involved in L2 performances. Chalhoub-Deville emphasizes that there is no one-size-fits-all rating scale (1995), since different aspects of the construct get emphasized depending on the speaking task, the raters who assess the speech, and other sources of systematic variance. Of course, the way the construct is operationalized in a rating instrument is constrained by our understanding of different dimensions of the construct and the way they are manifested at different proficiency levels (Iwashita, Brown, McNamara, & O'Hagan, 2008). A research priority, therefore, should be to more concretely define the constructs we claim to operationalize in our rating scales. This knowledge can then be extrapolated to operational L2 assessment contexts that view pronunciation as an integral part of the construct of L2 oral proficiency.

Endnotes

[1] An assumption of measuring an ability on a continuum is that there is an incremental increase in the unidimensional ability at each successive scale level.

[2] Flege and Fletcher (1992) measured listeners' accentedness ratings on a 256-point sliding scale. The position of the lever on the device indicated the score that was assigned.

[3] Raters were unaware that rating scale length was being manipulated in this study.

[4] We also ran separate 3-way ANOVAs for comprehensibility, accentedness, and fluency to test for interaction effects between the scale length, rater experience, and verbal protocol conditions. No significant interactions were detected.

Acknowledgments

This research was supported by Social Science and Humanities Research Council of Canada doctoral and postdoctoral fellowships awarded to the first and second authors respectively and by a Sir James Lougheed Award of Distinction awarded to the second author. An earlier version of this paper was presented at the annual meeting of the *American Association of Applied Linguistics* in Washington, DC (May, 2008) and at the *Language Testing Research Colloquium* in Denver, CO (March, 2009). The authors gratefully acknowledge Tracey Derwing and Murray Munro for sharing their speech elicitation and rater training materials, Carolyn Turner and Pavel Trofimovich for their valuable feedback on an earlier version of this manuscript, and our research assistants Marc Garellek, Daniel Clement, Joseph Hartfeil, and Monika Spak.

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy*. London: Macmillan.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*, 86-107.

Barnwell, D. (1989). Proficiency and the native speaker. *ADFL Bulletin*, *20*, 42-46.

Bendig, A. W. (1953). The reliability of self-ratings as a function of the amount of verbal anchoring and of the number of categories on the scale. *Journal of Applied Psychology, 37*, 38-41.

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

Brindley, G. (1991). Defining language ability: The criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing* (pp. 139-164). Singapore: SEAMEO.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks*. TOEFL Monograph 29. Princeton, NJ: Educational Testing Service.

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

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. Language Testing, *12*, 62-70.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

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

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*, 31-51.

Davidson, F. (1991). Statistical support for training in ESL composition rating. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 155-164). Norwood, NJ: Ablex.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*, 1-15.

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

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., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*, 665-679.

Derwing, T. M., Thomson, R. I., & Munro, M. J. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System, 34*, 183-193.

DeVelle, S. (2008). The revised IELTS pronunciation scale. *Research Notes, 34*, 36-38.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*, 125-144.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Flege, J. E., & Fletcher, K. L. (1992). Talker and listener effects on degree of perceived foreign accent. *Journal of the Acoustical Society of America, 91*, 370-389.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*, 208-238.

Garner, W. R. (1960). Rating scales, discriminability, and information transmission. *Psychological Review, 67*, 343-352.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning, 41*, 337-373.

Harding, L. (2008). Accent and academic listening assessment: A study of test-taker perceptions. *Melbourne Papers in Language Testing, 13*, 1-33.

Huck, S. W. (2004). *Reading statistics and research* (4th ed.). Boston: Pearson.

Isaacs, T., & Trofimovich, P. (in press). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of L2 speech. *Applied Psycholinguistics*, *32*, 113-140.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*, 24-49.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26*, 485-505.

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

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

Linacre, J. M. (1999). Category disordering vs. step (threshold) disordering. *Rasch Measurement Transactions, 13*, 675.

Linacre, J. M. (2005). A user's guide to FACETS: Rasch-model computer programs [Software manual]. Chicago: Winsteps.com.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? *Educational and Psychological Measurement, 31*, 657-674.

McKelvie, S. J. (1978). Graphic rating scales: How many categories? *British Journal of Psychology, 69*, 185-202.

McNamara, T. F., & Roever, C. (2006). *Language Testing: The social dimension*. Malden, MA: Blackwell.

Milanovic, M., Saville, N., Pollitt, A., & Cook, A. (1996). Developing rating scales for CASE: Theoretical concerns and analyses. In M. Milanovic, N. Saville & A.

Pollitt (Eds.), *Validation in language testing* (pp. 15-38). Clevedon, UK: Multilingual Matters.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Munro, M. J., & Derwing, T. M. (1994). Evaluations of foreign accent in extemporaneous and read material. *Language Testing, 11*, 254-266.

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

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28*, 113-131.

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

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*, 1-15.

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review, 65*, 395-412.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing, 14*, 157-184.

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

Stemler, S. E., & Tsai, J. (2008). Best practices in estimating interrater reliability. In J.

Osborne (Ed.), *Best practices in quantitative methods* (pp. 29-49). Thousand Oaks, CA: Sage.

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

Trofimovich, P., Gatbonton, E., & Segalowitz, N. (2007). A dynamic look at L2 phonological learning: Seeking processing explanations for implicational phenomena. *Studies in Second Language Acquisition, 29*, 407-448.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*. Cambridge: Cambridge University Press.

Upshur, J. A., & Turner, C. E. (1999). Systemic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing, 16*, 82-111.

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System, 36*, 69-84.

Connecting Text – Study 2 to Study 3

Studies 1 and 2 examined raters' cognitive abilities and L2 teaching experience as a function of their scalar judgments of L2 comprehensibility, accentedness, and fluency. Overall, these rater characteristics had little bearing on the numerical scores raters assigned. Qualitative group differences in experienced and novice raters' approach to the task in Study 2 were nuanced and revolved precisely around the criterion that was used to distinguish them—L2 teaching experience (e.g., access to ESL jargon, the use of L2 learners as a point of reference while rating).

Rating scale length was also systematically investigated in Study 2. No differences were detected in mean ratings assigned using 5-point and 9-point scales, overall score distributions were similar, and Cronbach's alpha values were high across experimental conditions. However, rater comments on rating scale use and preference coupled with the overlapping categories of the Rasch probability plots suggest that raters would benefit from additional guidance on how to interpret the focal construct, on the quality of speech that is characteristic at each scale level, and on salient criteria for making level distinctions.

The numerical scales used in Studies 1 and 2 do not provide much support to this effect. The generic way that the construct is articulated in these scales (i.e., it is unclear what different scale points "mean" in terms of performance features) does not answer the question "what is the construct?" with clarity or precision (Bachman, 2007, p. 41). While such Likert-type scales are useful for rank ordering L2 learners from a variety of L1 backgrounds on the focal construct, they would be limited in their utility in classroom settings, since the numerical values are not aligned with written descriptors beyond the broad range given at the scale anchors. Thus, the quality of feedback that could be given

to L2 learners on the basis of their performances when scores are recorded using numerical rating scales is obviously limiting.

In light of these shortcomings, the goal of Study 3 was to construct an empirically-based L2 comprehensibility scale that builds into the scoring rubrics the qualities of the L2 performance that appear to be most salient to raters. Fulcher's (1996) construction of a data-driven L2 fluency scale over a decade ago sets an important precedent for this study. Like fluency, comprehensibility has been defined in many ways in the literature (see definitional clarification in Study 3) and stands to benefit from clearer operationalization in rating scales. Given the importance of comprehensibility as a major construct in L2 pronunciation research (Derwing & Munro, 2009b), its congruence with communicative principles (Morley, 1994), and its role in contributing to communication breakdowns among interlocutors (Lindemann, 2006), it is, perhaps, surprising that no previous data-driven L2 comprehensibility scales have been developed. Comprehensibility is, therefore, the sole focus of Study 3, and the process and product described in the paper is the development of an empirically-based L2 comprehensibility scale.

The 9-point numerical comprehensibility scale in Studies 1 and 2, used by convention in L2 pronunciation research, is the benchmark against which the developed L2 comprehensibility scale will eventually need to be compared. It also serves as the starting point for rating scale development in Study 3, since it gives raters more scoring leeway (i.e., "rating room") than a 5-point scale, which some raters in Study 2 found too restrictive. However, on the basis of the finding in Study 2 that nine levels are difficult for raters to manage (see also Hamp-Lyons & Henning, 1991), it was decided *a priori* that the empirical L2 comprehensibility scale, in this early stage of development, would

consist of only three levels (low, intermediaxte, and high), and that this "crude" scale would be reworked and refined in future piloting and validation studies in view of the number of levels that raters are reliably able to distinguish. The descriptors reflect both the linguistic features encountered in the L2 performance samples (Fulcher, 2008), and raters' indications of influences on their judgments (Brown et al., 2005).

Notably, the L2 speaker and rater data in Study 3 were drawn from the same dataset used in Study 1. In the case of the rater data (n = 60), the music and non-music majors categories were collapsed, since no significant differences in mean L2 comprehensibility ratings were detected between these rater groups in Study 1. Thus, the focus here was on examining listeners' L2 comprehensibility judgments in general, and music major status did not need to be taken into account due to the null result.

In addition to these data sources, introspective report data from three ESL teachers were collected in Study 3 for data triangulation purposes. Because introspective reports (verbal protocols) proved an effective way of eliciting raters' perspectives in Study 2, they were incorporated into Study 3. However, rather than verbalizing their thoughts, the ESL teachers in Study 3 typed their impressions of the speech and salient influences on their judgments into a textbox directly below each rating scale in a word processing document. This written medium reflected the fact that the interest in Study 3 was not in rater cognitive processes while they conducted a problem-solving (i.e., rating) task *per se*, but rather in the factors that they attended to while rating, to inform the rating scale under construction.

Finally, there were suggestions in Study 2 that experienced raters are more adept in their descriptions of L2 speech than novice raters, due presumably to their greater exposure to linguistic terms and experience describing and evaluating learner

interlanguage. This finding informed the decision to recruit ESL teachers with graduate-level training and extensive teaching experience to provide the written reports in Study 3, since it was assumed that these raters would be better able to pinpoint the linguistic influences on their judgments than novice raters and, thus, provide more informative accounts for rating scale development. Further, because no quantitative group differences were detected between experienced and novice raters in Study 2, it was assumed that the overall ratings assigned by the teacher raters would not deviate much from those assigned by the 60 novice raters, but could supplement statistical analyses involving those ratings. In sum, Studies 1 and 2 were essential to the narrative that is presented in Study 3, both in terms of the rationale of the study, and informing the methodological decisions.

Chapter 4 – Study 3

Deconstructing L2 comprehensibility:

A data-driven approach to rating scale development

Talia Isaacs, McGill University

Pavel Trofimovich, Concordia University

Isaacs, T., & Trofimovich, P. (under revision). Deconstructing L2 comprehensibility: A data-driven approach to rating scale development. *Studies in Second Language Acquisition*.

Deconstructing L2 comprehensibility:

A data-driven approach to rating scale development

Abstract

Comprehensibility, or listeners' perceptions of how easily they understand second language (L2) speech, has been a major concept in L2 pronunciation research. Comprehensibility is congruent with the instructional goal of helping L2 learners achieve intelligible pronunciation, and is central to interlocutors' communicative success in real-world contexts. Comprehensibility has also been featured in oral proficiency scales for several high-stakes tests (e.g., TOEFL, IELTS). However, these scales tend to be limited by vague descriptions of comprehensible pronunciation or by conflating comprehensibility and accentedness, which are, in fact, partially independent dimensions. These shortcomings reflect the lack of empirical evidence about the linguistic aspects which influence listeners' judgments of L2 comprehensibility at different ability levels. In order to address this gap, a mixed-methods approach was used in the present study to develop an empirical L2 comprehensibility scale. First, the extemporaneous speech samples of 40 native French learners were analyzed using 19 quantitative speech measures, which included segmental, suprasegmental, temporal, grammatical, lexical, and discourse-level variables. These measures were then correlated with 60 native listeners' scalar judgments of the speakers' comprehensibility. Next, three experienced ESL teachers provided introspective reports to examine the linguistic aspects of speech that they attend to when judging L2 comprehensibility. By relating the results of these analyses, it was possible, first, to identify the L2 speech variables that distinguish

between L2 learners rated at different comprehensibility levels, and, second, to elaborate rating descriptors. Overall, lexical richness and fluency measures differentiated between low-level learners, grammatical and discourse-level measures differentiated between high-level learners, and word stress errors discriminated between learners at all levels.

Deconstructing L2 Comprehensibility:

A Data-Driven Approach to Rating Scale Development

Introduction

For the past several decades, communicative approaches to language teaching have been a dominant force in a variety of instructional settings (Celce-Murcia, Brinton, & Goodwin, 1996; Elliott, 1997; Gor & Vatz, 2009). Because nontarget pronunciation can lead to unintelligible speech and communication breakdowns (Fayer & Krasinski, 1987; Jenkins, 2000; Munro & Derwing, 1999), second language (L2) pronunciation would seem to play a central role in pedagogical approaches emphasizing interactions with interlocutors in the target language. However, Kelly dubbed pronunciation "the Cinderella of language teaching" as early as 1969 (p. 87), and pronunciation has since been characterized as suffering from "the 'Cinderella syndrome'—kept behind doors and out of sight" (Celce-Murcia et al., 1996, p. 323).

One reason for the exclusion of pronunciation from communicative L2 teaching comes from the belief that an overt focus on pronunciation is ineffective and even extraneous to helping learners achieve communicative competence (e.g., Krashen, 1981; Terrell, 1989). Morley (1991) counters this view by arguing that "intelligible pronunciation is an essential component of communicative competence" (p. 488) and that "ignoring students' pronunciation needs is an abrogation of professional responsibility" (p.489), since poor pronunciation can be professionally and socially disadvantageous to L2 speakers (Lippi-Green, 1997; Munro, 2003). There is also evidence that adult L2 learners with "fossilized" pronunciation do, in fact, benefit from explicit pronunciation

instruction (e.g., Couper, 2006; Derwing, Munro, & Wiebe, 1998) and that a focus on pronunciation can be embedded in genuinely communicative activities (Trofimovich & Gatbonton, 2006).

Repercussions of the marginalization of pronunciation are still being felt in classroom practice. One area where classroom teachers have received virtually no support is in the provision of formative assessment tools, such as rating scales, to help describe and benchmark learners' performance. The main issue here is that not enough is known about major constructs in L2 pronunciation (e.g., intelligibility, comprehensibility, accentedness) to adequately operationalize them in L2 scales for assessment purposes. Therefore, the goal of the present study is to fill this research gap by examining the linguistic aspects of L2 speech that most contribute to comprehensibility (listeners' perception of how easily they understand L2 speech), for the purpose of empirically deriving a descriptive L2 comprehensibility scale.

Why a Focus on Comprehensibility?

Few L2 researchers and practitioners would disagree that intelligibility is the appropriate goal for L2 pronunciation instruction. The chief reason for this is that, in most situations of L2 use, what really counts is L2 speakers' ability to be understood, rather than, for example, the quality or nativelikeness of their accent (Derwing & Munro, 1997; Jenkins, 2000; Pennington & Richards, 1986). This raises the important question of why comprehensibility, rather than intelligibility, is targeted here for the development of a pronunciation scale. Levis' (2006) distinction between broad and narrow definitions of intelligibility described in Paper 3 is of relevance. In its narrow sense, intelligibility is defined as listeners' *actual* understanding of L2 speech (Munro & Derwing, 1999). It is typically measured as the proportion of words that a listener is accurately able to

decipher, as shown, for example, through orthographic transcriptions. In its broad sense, intelligibility refers more generally to a listener's ability to understand the speech and "is not usually distinguished from closely related terms such as comprehensibility" (Levis, 2006, p. 252).Comprehensibility is typically defined as listeners' *perceptions* of understanding and is measured through listeners' scalar ratings of how easily they understand speech (Munro & Derwing, 1999). In the context of L2 tests, several oral proficiency scales make use of the term "intelligibility" (e.g., TOEFL, IELTS). However, in all cases, "intelligibility" is measured in terms of listeners' subjective judgments obtained through scalar ratings, which suggests that it is, in fact, "comprehensibility" that is used as a criterion in these assessment tools. The construct of comprehensibility in the present study falls under Levis' broad sense of intelligibility and thus reflects a typical approach to assessing intelligibility in oral proficiency scales.

Comprehensibility in Theoretical Models

Theoretical models of communicative competence and communicative language ability (e.g., Bachman & Palmer, 1996; Canale & Swain, 1980) often serve as the basis for rating scale development (Luoma, 2004). However, the role of pronunciation (and comprehensibility in particular) has not been clearly defined in these models. The last major model to deal extensively with pronunciation, Lado's (1961) skills-and-components model, distinguishes skills of speaking, listening, reading, and writing from components of grammar, vocabulary, and phonology/graphology. Although Lado devotes separate chapters to testing L2 learners' perception and production of individual sounds, stress, intonation and offers guidelines for item construction and test administration with decontextualized examples, his model does not clarify the relationship between skills and components (Bachman, 1990) and is generally considered outdated, given expanded

notions of communicative ability and advancements in language testing (Bachman, 2000). In their influential model of communicative competence, Canale and Swain (1980) list knowledge of the rules of phonology as part of grammatical competence, but do not provide a definition of phonology or clarify its applicability to L2 learners. Similarly, in Bachman's communicative language ability framework (1990) and its further development in Bachman and Palmer (1996), reference to "phonology/graphology" appears to be a carryover from Lado, again with no definition of either construct. Therefore, the role of pronunciation in models of communicative competence and communicative language ability needs to be defined more clearly so that these models can better inform the development of communicatively-oriented rating scales.

Comprehensibility in L2 Assessment Instruments

There are several shortcomings in the way that pronunciation, and comprehensibility in particular, have been modeled in existing L2 speaking scales. The treatment of pronunciation in oral proficiency scales is often inconsistent, if included at all. Levis, for example, describes the pronunciation component of the American Council of the Teaching of Foreign Languages (ACTFL) Oral Proficiency Guidelines (Breiner-Sanders, Lowe, Miles, & Swender, 2000) as a "haphazard collection of descriptors" and "strikingly random in describing how pronunciation contributes to speaking proficiency" (2006, p. 245). In other scales, such as the holistic and analytic Common European Framework of Reference (CEFR) scales, pronunciation is omitted altogether from the descriptors of benchmark levels (Council of Europe, 2001; see also North, 2000).

Even when included, pronunciation descriptions are often too vague to delineate a coherent construct. For example, Band 4 in the publicly available version of the IELTS Speaking Band Descriptors reads: "uses a limited range of pronunciation features;

attempts to control features but lapses are frequent; mispronunciations are frequent and cause some difficulty for the listener" (British Council, IDP: IELTS Australia, & UCLES, n.d.). Similarly, the TOEFL iBT Integrated Speaking Rubrics link "intelligibility" with "pronunciation," "intonation," and "pacing" (ETS, 2005). Notably, the descriptors in both scales are vague and do not identify the types of errors that lead to listener difficulty, since some errors could be more detrimental to comprehensibility than others (e.g., Munro & Derwing, 2006). The use of the term "pronunciation" is likewise not consistent across these scales. Whether the term refers solely to segmental features (i.e., errors involving individual sounds) or also encompasses other aspects of speech, including suprasegmental features (e.g., word stress, rhythm, intonation), needs to be clearly spelled out to facilitate the interpretation of scale descriptors for both raters and test users.

Relativistic wording in rating scales offers even less clarity about the construct being measured. The scale bands in Morley's (1994) Speech Intelligibility/Communicability Index, for example, make reference to "fully," "largely" or "reasonably intelligible" and "basically" or "largely unintelligible" speech. Further, the numerical comprehensibility scales used in L2 research normally range from "extremely difficult to understand" to "extremely easy to understand" (e.g., Derwing & Munro, 1997), with no further definition provided to raters (Derwing, Munro, & Thomson, 2008). Notably, the numerical scales are low-stakes scales (i.e., used for research purposes) unlike the scales associated with the standardized tests cited above. While the lack of detailed descriptors in these scales means that they can be used with learners from all L1 backgrounds, raters are unlikely to assign a common meaning to the values that designate different levels of the scale for any given set of speech samples to be rated in the absence of additional guidance (Isaacs & Thomson, 2010). Although interrater reliability

(typically estimated through intraclass correlations) is almost universally high using this rating procedure (e.g., Isaacs & Trofimovich, 2011; Derwing et al., 2008), raters need both a clear definition of comprehensibility and detailed information on its key aspects.

Current L2 speaking scales also frequently conflate the dimensions of comprehensibility and accentedness (Harding, in press). Morley's (1994) Speech Intelligibility/Communicability Index, for example, equates incremental increases in comprehensibility with incremental decreases in accentedness up to the highest level, where "near-native" speech is accompanied by a "virtually nonexistent" accent (p. 77). Similarly, the highest level of Cambridge ESOL's Common Scale for Speaking links "native-like" control of "many features" with easily understandable pronunciation (UCLES, 2008, p. 70), and comprehensibility and accentedness are grouped together in many band descriptors of the CEFR scale of Phonological Control, one of several CEFR scales on distinct aspects of competence (Council of Europe, 2001). Part of the reason for the juxtaposition of accentedness and comprehensibility in rating scales is that, apart from work on accent, there is a dearth of empirical research describing the qualities of comprehensible speech (for rare exceptions, see Fayer & Krasinski, 1987, and Varonis & Gass, 1982, both reviewed below). The critical point here is that accentendess does not *necessarily* lead to poor comprehensibility or communication breakdowns, but tends to be overemphasized due to its perceptual salience (Derwing & Munro, 2009).

Empirical Development and Validation of L2 Rating Scales

In light of these shortcomings, there is an urgent need for an empirically-derived rating scale which can describe the factors that might influence listeners' judgments of L2 speech at different levels of comprehensibility. Although comprehensibility is important for both high-stakes, rater-mediated speaking assessments and for informal judgments of

L2 speech in real-world interactions, to our knowledge, no study has focused on the development of such a scale using a data-driven approach. Nevertheless, previous research has influenced the development of a L2 comprehensibility scale in the present study.

One such example is Fulcher's (1996) work on L2 fluency. Fulcher argued that test developers did not have an adequate definition of fluency that could be operationalized in rating scales for assessment purposes. His approach was to use grounded theory to generate a "thick description" of fluency at different ability levels. This involved coding 21 ELTS interview transcriptions (a precursor to the IELTS) to generate explanatory fluency categories (e.g., hesitations due to content planning, lexical access, etc.). The coded categories were then tallied and cross-validated using discriminant analysis. Results showed that the researcher-generated categories discriminated well among test-takers, accurately predicting the ELTS band score placement for all but one test-taker. Finally, Fulcher elaborated a set of detailed fluency descriptors by focusing on those fluency categories that had provided the strongest level distinctions. The present study extends Fulcher's work by consulting raters directly about influences on their judgments through the use of a think-aloud procedure.

In a more recent study on the validation of TOEFL iBT speaking scales, Brown, Iwashita, and McNamara (2005) found close correspondence between the aspects of speaking proficiency that raters attended to without the guidance of a rating instrument and several quantitative measures used to analyze test-taker discourse. In a follow-up study, Iwashita, Brown, McNamara, and O'Hagan (2008) examined which of these measures, grouped into the broad categories of "linguistic resources," "phonology," and "fluency," distinguished between five levels of L2 speaking proficiency. Results showed

that measures from each category were captured in raters' score assignments, which implies that raters weigh multiple factors when assessing L2 oral proficiency. Iwashita et al. acknowledged the absence of discourse-level measures in their analyses as a limitation. In addition, the "pronunciation," "intonation," and "rhythm" measures in the phonology category involved impressionistic judgments by two trained phoneticians rather than more objective measures. The present study builds on Iwashita et al.'s research by examining performance in each of their overarching categories while also including discourse-level measures and finer-grained, more objective measures of phonology.

The Current Study

Clearly, there is a need to better understand the role of pronunciation within the broader realm of L2 oral proficiency and communicative competence. The starting point in the current study was to "unpack" comprehensibility, a major construct in the L2 pronunciation literature. Examining the factors that influence listeners' L2 comprehensibility judgments is ecologically valid, as listeners' impressions of the effort needed to understand L2 speech likely shape their real-world interactions with L2 interlocutors. Identifying the linguistic variables that contribute to L2 comprehensibility at different ability levels could also inform rater training in both low-stakes research settings and in high-stakes assessment contexts. Finally, knowledge of the aspects of speech that contribute to comprehensibility could help teachers set instructional targets, integrate pronunciation with the teaching of other skills, and inform formative assessment practices in the L2 classroom.

A sequential mixed-methods design was used to develop the data-driven comprehensibility scale in the present study (Creswell & Plano-Clark, 2007), where

evidence from earlier phases cumulatively informed subsequent phases. The first source of evidence was based on a quantitative analysis of speech measures associated with listeners' comprehensibility judgments. For this analysis, the speech of 40 French learners of English was presented to 60 listeners for comprehensibility rating. The same speech samples were then analyzed for 19 linguistic measures (including aspects of phonological accuracy, grammatical accuracy, lexical richness, story cohesion) to determine which measures were related to comprehensibility ratings. The second source of evidence was based on listeners' qualitative reports on the aspects of speech that they attended to when assigning comprehensibility ratings. For this analysis, a coding scheme was developed based on three experienced ESL teachers' detailed comments about the aspects of speech they focused on while rating. These descriptive comments were later "quantitized" by tabulating frequency counts of coded categories (Teddlie & Tashakkori, 2009). By combining the analysis of learner discourse with the listener reports, it was possible to identify the measures that differentiated between L2 learners at different levels of comprehensibility. The outcome was a scale featuring elaborated L2 comprehensibility descriptors.

Method

L2 Speakers

The speakers were 40 Francophones (13 males, 27 females) from a predominantly French speaking area of Quebec, Canada (M_{age} = 35.6, range = 28–61) who had participated in an earlier study on L2 phonological learning (Trofimovich, Gatbonton, & Segalowitz, 2007). With the exception of two early French-English bilinguals, the speakers had been exposed to English in 45-min weekly ESL classes in primary school and had received up to three hours per week of subsequent ESL instruction. At the time of

the study, the speakers estimated using English only 20% of the time on average, although to varying degrees (0–70%). Their self-reported English speaking and listening ability was also variable, spanning the entire range of the 9-point scale (*1 = extremely poor*, *9 = extremely proficient*). Overall, the speakers represented different ability levels (particularly with respect to their ability to speak English), from beginning to advanced (see Trofimovich et al.).

All speakers were recorded telling a picture story in English. The recordings took place in a quiet office using a Plantronics (DSP-300) microphone connected to a desktop computer. The eight-frame picture story used to elicit the speech featured two strangers who bumped into each other on a busy street corner. They dropped the identical suitcases they were carrying, only to later discover that they had accidently retrieved the wrong suitcase (Derwing et al., 2008). After normalizing the speech samples for peak amplitude and removing any initial dysfluencies (e.g., false starts, hesitations), the beginning of each narrative (23–36 s in duration) was excised from the recording and randomized for presentation to raters. In preparation for data analysis, the speech samples were transcribed and transcription accuracy was verified by a second transcriber.

L2 Speech Measures

The construct of comprehensibility has primarily been associated with research on L2 pronunciation. However, it is unlikely that, given a scale with the endpoint descriptors "very easy/difficult to understand," raters focus solely on phonological aspects of speech. In an attempt to capture as many variables as raters possibly use to arrive at their comprehensibility judgments, four categories of measures were considered. Three categories were the same as those used by Iwashita et al. (2008) in their study on L2 oral proficiency: *phonology*, which included segmental and suprasegmental measures; *fluency*,

which involved temporal measures; and *linguistic resources*, which comprised grammatical and lexical measures. The fourth category called *discourse* was added to capture speakers' storytelling strategies and use of cohesive devices, since these variables could influence raters' judgments if they interpret comprehensibility to mean understanding the message or story rather than each individual word (Isaacs & Thomson, 2009).

Phonology

Six measures were included in this category: two at the level of individual segments (vowels and consonants) and syllables, and four at the level of words and phrases.

(1) Segmental error ratio: defined as the number of phonemic (e.g., *fun* spoken as *fan*) and phonetic (e.g., the [t^h] in *time* spoken without aspiration) substitutions divided by the total number of segments articulated.

(2) Syllable structure error ratio: defined as the total number of vowel and consonant epenthesis (insertion) and elision (deletion) errors (e.g., *they* with an epenthetic schwa added at the end; *apologize* with schwa deletion at the beginning) over the total number of syllables articulated.

(3) Word stress error ratio: defined as the total number of instances of misplaced or missing major word stress in polysyllabic words (e.g., *BUIL-ding* spoken as *buil-DING*; *SKY-scra-per* spoken as *sky-scra-PER*) divided by the total number of polysyllabic words produced. The first three measures were drawn from a study by Anderson-Hsieh, Johnson, and Koehler (1992), who found a relationship between these measures and ratings of "intelligible speech" and accent combined in a single scale.

(4) Rhythm ratio: defined as the number of correctly reduced syllables over the total number of obligatory vowel reduction contexts in both polysyllabic words and function words (e.g., *in a CI-ty there was TWO PEO-ple* contains 6 obligatory contexts, all in lowercase letters; the speaker pronounced "people" as *peo-PLE* and, thus, produced 5 correct vowel reductions). This measure was designed to capture the stress-timed nature of English rhythm (Deterding, 2001).

(5) Phrase final pitch ratio: defined as the number of correct pitch patterns at the end of phrases (i.e., at syntactic boundaries) over the total number of appropriate pause locations at the end of phrases, where pitch patterns are expected (e.g., the sentence *it's a nice sunny afternoon in Montreal* [level tone] *when Bob and Margaret are walking down the street about to turn a corner* [falling tone] has two correct pitch patterns). This intonation measure was influenced by Wennerstrom's (2001) boundary tone measure but was judged auditorily rather than through instrumental analysis (Pickering, 2001).

(6) Pitch range: expressed as the difference between the highest and lowest fundamental frequency (F0) values measured in a pitch tracker across the speech sample, with spurious values omitted (see Cooper & Sorensen, 1981). Examples of pitch range, used here as a measure of how monotonous a person sounds, were 99.7 Hz (100.8–200.5) for a male speaker and 220.8 Hz (139.2–360.0) for a female speaker. F0 values were measured using *Praat* speech analysis software (Boersma & Weenink, 2010). This measure, based on Kang's (2010) overall pitch range measure, was influenced by Wennerstrom's (2001) notion of paratones, or pitch expansion to signal topic shift.

Fluency

Derwing, Rossiter, Munro, and Thomson's (2004) finding that listeners' scalar L2 comprehensibility judgments are statistically associated with temporal (fluency) measures prompted the analysis of seven temporal measures in the present study. For measures based on pause duration, the cutoff value for measuring pauses was set at 400 ms following Derwing et al. and Riggenbach (1991).

(7) Total number of filled pauses: defined as nonlexical pauses, such as uh and um (e.g., *it's a nice day in uh uh* [two filled pauses] *New York*).

(8) Total number of unfilled pauses: defined as silent pauses (e.g., *One day* [unfilled pause] *I was appointed* [unfilled pause] *to attend a meeting in New York City*). Filled and unfilled pauses as indexes of fluency were counted separately following Lennon (1990).

(9) Pause error ratio: defined as the number of inappropriately produced filled and unfilled pauses (i.e., inside clauses and not at syntactic boundaries, where pauses would be expected), divided by the total number of pauses produced overall (e.g., *They uh* [filled pause] *continue* [unfilled pause] *to walk to the* [unfilled pause] *work*).

(10) Repetition/self-correction ratio: defined as the sum of all immediately repeated and self-corrected words (e.g., *I I* [repeated word] *see uh buildings a a* [repeated word] *lot of buildings with uh in* [self-corrected word] *in* [repeated word] *a big city*) over the total number of words produced. Self-repetitions and corrections were pooled due to relatively few instances of self-corrections in the speech samples. Repetitions and self-corrections embedded in longer phrases (e.g., *buildings… a lot of buildings*) were not considered as such. This measure was

included to estimate possible detrimental effects of "stuttering" on listener comprehensibility (Isaacs & Thomson, 2009).

(11) Pruned syllables per second: defined as the total number of syllables produced excluding dysfluencies (e.g., filled pauses, repetitions, self-corrections, false starts), calculated over the total duration of the speech sample. Derwing et al. (2004) found this temporal measure to be the strongest predictor of raters' global L2 fluency judgments, and fluency and comprehensibility were, in turn, strongly correlated.

(12) Mean length of run (MLR): defined as the average number of syllables between two adjacent filled or unfilled pauses (e.g., Riggenbach, 1991).

Linguistic Resources

Because comprehensibility has mostly been studied in the context of L2 pronunciation research, investigations of other influences on comprehensibility, especially those that extend beyond phonological and temporal variables, have been limited. Two exceptions are early studies by Varonis and Gass (1982), who found that ungrammatical sentences have an overall negative effect on comprehensibility, and Fayer and Krasinski (1987), who showed that comprehensibility judgments were related to pooled mean ratings of grammar, pronunciation, intonation, and word choice.[1] To examine possible detrimental effects of grammatical and lexical errors on listener comprehension, one grammatical accuracy measure and three lexical measures were included.

(13) Grammatical accuracy: defined as the number of words with at least one morphosyntactic error divided by the total word count. Morphosyntactic errors were errors in sentence structure, morphology, or syntax, including word order

errors (e.g., *they falled on the floor and exchanged your suitcase* contained one verb conjugation error and one pronoun error). This measure is similar to Foster and Skehan's (1996) and Skehan and Foster's (1999) global accuracy measure, which was sensitive to differences in L2 oral performance as a function of task characteristics (e.g., planning time).[2] The measure of grammatical accuracy, as defined here, was conservative in the sense that no multiple morphosyntactic errors per word were counted (e.g., *there's a little house where live a woman* contained both a verb tense error and a word order error associated with the word *live*, but only one error was counted). This was done in order to control for extreme cases of variability in individual speakers' error counts.

(14) Lexical error ratio: defined as the number of incorrectly used lexical expressions, including phonetically similar but semantically inappropriate words (e.g., *above to arrive* instead of *about to arrive*), false cognates (e.g., *circulation* instead of *traffic*), imprecise vocabulary choice (e.g., *carrying bags* instead of *carrying suitcases*), incorrectly used lexical expressions (e.g., *walkside* instead of *sidewalk*) and L1 intrusions (e.g., *ah mon dieu les temps en plus*), over the total number of words produced (see Swan, 1997, for a discussion of lexical errors due to lexical transfer).

(15) Token frequency: defined as the total number of words produced (Laufer & Nation, 1995).

(16) Type frequency: defined as the total number of unique words produced. Types and tokens were calculated separately using the online *Vocabprofile* program (Cobb, 2000).[3] Because type and token frequencies are sensitive to sample length, both measures were normalized by dividing the frequencies by the total duration of the

sample. The resulting measures thus represented the frequencies of word tokens and types per unit of time.

Discourse

Because listeners may also rely on speakers' storytelling strategies and attend to the discourse structure of speakers' narratives in making comprehensibility judgments (see Isaacs & Thomson, 2009, for qualitative evidence), three discourse-level measures were examined.

(17) Story cohesion: defined as the number of adverbials used as cohesive devices (Martin & Rose, 2003). These devices (e.g., *suddenly*, *but*, *hopefully*) help situate the story by establishing links between storytelling elements, propelling the storyline forward, or revealing the storyteller's attitude.

(18) Story breadth: defined as the number of distinct propositions or storytelling elements used by a speaker. Propositions were identified using Stein and Glenn's (1979) scheme, which includes such categories as setting (e.g., *the story is beginning in Manhattan*), initiating event (e.g., *I rush at the office this morning with my briefcase*), attempt (e.g., *so they banged into each other*), direct consequence (e.g., *they went to took their luggage but they took the wrong one*), and reaction (e.g., *the two person are confuse*).

(19) Story depth: defined as the number of different proposition categories used by a speaker (e.g., setting, attempt, reaction). A L2 speaker whose story dealt exclusively with the setting, for example, would receive a lower score on this measure than a speaker who briefly set the scene, then described the events and consequences. Because, as with lexical variables, discourse measures are sensitive to sample length, all discourse measures were normalized by dividing the

frequencies by the total duration of the sample. The resulting measures thus represented the frequencies of cohesive devices, propositions (story breadth), and proposition categories (story depth) per unit of time.

Following initial coding by a trained coder, another trained coder recoded 40% of the speech samples for each of the 19 measures in order to establish intercoder reliability. Intraclass correlations for each measure were .90 or higher, with the exception of lexical error ratio (.85), revealing overall high intercoder agreement.

Phase One: Quantitative Data

The purpose of the first phase was to examine which of the 19 speech measures were most strongly related to raters' L2 comprehensibility judgments. This was viewed as the first step to inform the construction of the eventual L2 comprehensibility scale.

Method

Raters and Rating Procedure

L2 comprehensibility judgments were obtained from 60 raters. The raters were native English-speaking undergraduate students (26 males, 34 females) majoring in a variety of non-linguistic disciplines (e.g., physiology, music, sociology, biochemistry) at an English-medium university in Montreal, Canada. The raters (M_{age} = 20.7, range = 19–25) reported growing up in monolingual homes in Canada (29) and the United States (31), estimated speaking and listening to English over 90% of the time daily, and rated their French (L2) speaking and listening ability at a low-intermediate level (3.4) on a 9-point scale (*1 = extremely poor*, *9 = extremely proficient*). All raters reported having normal hearing. Because the raters lacked L2 teaching experience and specific language training, they were considered "novice raters."

The speech samples were presented to individual raters in a quiet room in a self-paced task via a Koss R/80 headset connected to a desktop computer. After listening to each picture story in randomized order, the raters assigned comprehensibility scores on separate 9-point scales (*1 = hard to understand*, *9 = easy to understand*). As part of an larger study (Isaacs & Trofimovich, in press), the raters also evaluated the speech samples for accentedness and fluency, and were assessed on several cognitive variables (e.g., attention control).

Results

Intraclass correlations were calculated first to examine whether the novice raters were internally consistent in their ratings. A coefficient of .97 suggested that the raters were consistent in their judgments. Pearson correlations were then computed to examine the strength of the relationship between the mean L2 comprehensibility ratings, averaged for each speaker across the 60 raters, and the 19 analyzed speech measures. Table 1 shows that strong correlations ($r > .70$) were found for several measures in each of the conceptual categories of phonology (word stress error ratio, rhythm ratio), fluency (MLR), linguistic resources (type frequency, token frequency), and discourse (story breadth). Moderate correlations ($r > .40$) were revealed for 9 of the 13 remaining measures, with only one measure showing no relationship with comprehensibility (pitch range). This suggests that L2 comprehensibility ratings are related to a wide range of variables that clearly are not restricted to the domains of phonology and fluency.

Table 1. *Pearson Correlation Coefficients Between L2 Speech Measures and 60 Novice Raters' Scalar Judgments of L2 Comprehensibility*

Speech measure	Correlation
Type frequency	.78**
Token frequency	.77**
Word stress error ratio	–.76**
Rhythm ratio	.74**
Mean length of run	.71**
Story breadth	.71**
Grammatical accuracy	–.63**
Pause error ratio	–.58**
Phrase final pitch ratio	.57**
Repetition/self-correction ratio	–.57**
Segmental error ratio	–.54**
Lexical error ratio	–.52**
Story cohesion	.50**
Total filled pauses	–.45**
Story depth	.42**
Syllable structure error ratio	–.37*
Pruned syllables per second	.35*
Total unfilled pauses	–.32*
Pitch range	–.07

Note. *$p < .05$, **$p < .01$, two-tailed.

Phase Two: Qualitative Data

The purpose of the second phase was to generate listener input into the aspects of L2 speech that they consider when judging comprehensibility. Building rater perceptions into the scale development process was necessary to ensure that the eventual scale reflected not simply the most statistically robust measures, but also the most salient criteria that the intended users of the scale (i.e., raters and especially teachers) attend to when making comprehensibility level distinctions.

Method

Teachers

Following previous research that has drawn on experienced teachers' perspectives in the development and validation of rating scales (e.g., North, 2000; Upshur & Turner, 1999), three native English-speaking ESL teachers (1 male, 2 female) with 10–12 years of classroom experience were consulted. Originally from Western Canada, the teachers had moved to Montreal as adults, with 8 to 24 years of residency. All teachers were teaching Francophone learners of English at the time of the study but estimated speaking and listening to French less than 20% of the time. They had all taken graduate-level TESL courses; however, none had received training in L2 assessment or phonetics/phonology for teachers. Nonetheless, they were charged with classroom assessment responsibilities, and therefore came from a population who could potentially benefit from access to a user-friendly assessment tool for L2 comprehensibility.

An additional reason for examining experienced teachers' impressions of L2 comprehensibility is that a previous study by Isaacs & Thomson (2009) suggested that teacher raters were better able to articulate linguistic influences on their judgments in the absence of rating guidelines than novice raters, who tended to describe only a small set of default features in learners' speech (e.g., pausing, speech rate). Therefore, it was thought that experienced teachers would be more able to identify a fuller range of aspects of speech that they consider when scoring comprehensibility than novice raters, who may have less clearly developed internal criteria for L2 oral assessments or may lack the vocabulary for expressing their thoughts.

Ratings and Written Reports

The teachers were probed about their impressions of the speech and influences on

their ratings in individual self-paced sessions that did not exceed two hours. The teachers first familiarized themselves with the speech elicitation task (picture narrative) and completed several practice ratings. They then listened to the 40 speech samples in randomized order using a Koss R/80 headset. In order to provide initial standardization, comprehensibility was defined as "how easy the speaker is to understand." When the teachers were ready to score each speech sample, following multiple listenings if necessary, they paused the recording and typed their comprehensibility rating into a preformatted word processing document using the 9-point comprehensibility scale described above. Below each rating scale, the teachers then related the aspects of the speech that they attended to when scoring. Finally, at the end of the session, the teachers summarized their listening and rating experience in a follow-up questionnaire with space for open-ended comments. The teachers were specifically asked whether they had interpreted comprehensibility to mean comprehensibility of the individual words, comprehensibility of the story or message, or had adopted a different interpretation.

Results

The multiple sources of teacher data were initially analyzed separately, then combined to strengthen the interpretation of the findings. The rating data were first submitted to intraclass correlations to examine the teachers' scoring consistency, both with their peers and with the novice raters. Then, the written reports for each speech sample were coded as a function of L2 comprehensibility level. Finally, teachers' questionnaire comments about their interpretation of comprehensibility were used to clarify other sources of evidence.

Intraclass Correlations

The intraclass correlations for comprehensibility scores assigned by the three ESL

teachers (henceforth, T1, T2, and T3) showed that the agreement between T1 and T2 was relatively high (.81). However, the agreement between each of these teachers and T3 was lower (.62 and .66 respectively), revealing a poorer scoring consensus when T3 was involved. At least some of this divergence may reflect differences in teachers' understanding of the construct being measured. Whereas T1 and T2 interpreted comprehensibility as the listener's ability to understand the L2 speaker's story or message, T3's interpretation centered on the listener's ability to decipher the speaker's individual words. These differing perspectives suggest that comprehensibility may need to be defined more precisely in L2 research and assessment contexts than simply "ease of understanding", in order to support a more unitary interpretation for construct validity reasons.

Intraclass correlations between T1, T2, and T3 and pooled L2 comprehensibility ratings of the 60 novice raters yielded coefficients of .90, .88, and .80 respectively, which suggested that ratings pooled over a large group tend to average out individual raters' idiosyncrasies. Because the novice raters, compared to the teachers, showed a higher degree of concordance in their ratings, the 40 L2 speakers were rank-ordered by the novice raters' mean comprehensibility scores. The speakers were then classified into low (n = 13), intermediate (n = 13), and high (n = 14) L2 comprehensibility groups, so that the aspects of the speech that the teachers considered in their ratings could be examined as a function of speaker comprehensibility level.

Analysis of Written Reports

For the analysis of the written reports, a 10-category coding scheme was developed, with the analyzed speech measures from the previous phase serving as the starting point. The challenge was to generate a system in which the categories were

narrow enough that meaningful distinctions at different comprehensibility levels could be captured, but not so fine-grained that it would be difficult for another coder to reliably apply the categories. For example, the overlapping categories of "L1 intrusions," "L1-influenced lexical items," and "odd lexical choice," which had a clear conceptual link to the error types examined under the quantitative speech measure "lexical error ratio," were initially coded separately but later merged under the broad category of "vocabulary." When the coding was completed, 40% of the data were recoded by a second coder blind to the purposes of the study. Exact agreement was obtained for 95% of the observations, indicating high intercoder agreement. In instances where the coding was inconsistent, consensus was achieved through substantive discussion.

Frequencies of the coded categories by speaker comprehensibility level are shown in Table 2. Although each teacher emphasized different aspects of speech, taken together, coverage of the overarching conceptual categories used to group the quantitative speech measures (phonology, fluency, linguistic resources, discourse) was achieved. All three teachers commented on the coded categories of grammar, vocabulary, and fluency (see the first three rows in Table 2). The trend was that the number of comments for these categories was highest for the low-comprehensibility group and decreased at each subsequent level, although there appeared to be a leveling off between intermediate- and high-level groups for vocabulary.

Table 2. *Frequency of Coded Categories from Teacher Reports Grouped by L2 Speaker Comprehensibility Level*

Coded category	Speaker comprehensibility level			Total comments
	Low	Intermediate	High	
Grammar	22	14	9	45 (T1,T2,T3)
Vocabulary	17	11	10	38 (T1,T2,T3)
Fluency	14	9	3	29 (T1,T2,T3)
Inadequate words or information produced	6	-	-	6 (T1, T3)
Storytelling elements and cohesion	6	8	12	26 (T1)
Accent/pronunciation (general comment)	9	11	-	20 (T1, T3)
Word stress	2	4	-	6 (T3)
Intonation	2	2	-	4 (T3)
Need to be a teacher, know the context, or have exposure to French to understand	14	9	1	29 (T2)
Any listener can understand regardless of background	1	1	6	6 (T2)

Note. Comprehensibility level categorizations are based on novice raters' mean score assignments.

Grammar was the category with the highest number of observations overall. Most comments tended to be generic, although T1 and T2 pinpointed verb tense errors and, less frequently, pronoun and preposition errors in low- and intermediate-level speakers. Of the nine coded grammar comments at the high-comprehensibility level, seven were either positive in nature or, in T3's words, revealed "no grammar errors to distract," in contrast to the lower levels, where T3 often explicitly identified grammar as contributing to comprehension difficulties in conjunction with other aspects of speech. The vocabulary category, which had the second highest number of net observations, encompassed

instances of imprecise or L1-influenced vocabulary, odd lexical choice (e.g., "holding" a suitcase instead of carrying a suitcase), the use of phonetically similar but semantically inappropriate words (e.g., "crushed" for crashed), and in the case of low-comprehensibility learners only, French L1 intrusions. Similarly, for the fluency category, teachers commented on pauses, hesitations, repeated words, self-corrections, and pacing (e.g., representative comments included, "pace was slow," "lack of fluidity," "hesitations, corrections and repetition also delay understanding of message," etc.). These comments appeared to have counterparts in the analyzed quantitative speech measures. Reference to segmental errors, on the other hand, was conspicuously absent from all teacher reports, although T3 referred to "accent" and "pronunciation" in broad terms for low- and intermediate-level speakers.

There was less consistency across the teachers in other coded categories. For example, only T1 referred to discourse measures. Comments about storytelling elements and cohesion were pooled together due to their co-occurrence in T1's remarks (e.g., "no continuity to the story," "random images with no glue"). For speakers in the low-comprehensibility group, T1 often reported that he had "no idea what the story was about." Conversely, speakers at the high end of the spectrum evoked either positive comments (e.g., "good description of the weather and details of the first scene") or comments about the lack of story details (e.g., "doesn't give enough detail where needed like mentioning the people on the sidewalk"). T3 was also the only teacher to mention "syllable/word stress" and "intonation," although without providing examples. In fact, her strategy was to construct her own basic descriptor and then slightly modify it for the individual L2 speaker being rated. For seven low- and nine intermediate-level speakers, her description followed the formula, "(relatively) easy to understand in terms of

pronunciation," with a list of criteria that "(slightly) contributed to difficulties in comprehensibility," depending on which were applicable (e.g., grammar, hesitation, intonation etc.). In contrast, all high-comprehensibility speakers were either "perfectly" or "completely comprehensible" in her view.

T2 was also distinct from the other teachers, specifically in her overall orientation to rating. Her interpretation of comprehensibility strongly revolved around her assumption that listeners' knowledge of the context (i.e., picture story content), familiarity with the speakers' L1 (French), and ESL teaching experience would likely have a facilitative effect on their ability to understand the speech, whereas listeners without recourse to these factors may not be able to compensate for gaps in their understanding. The frequency counts of T2's comments in the bottom two rows of Table 2 show that the listener's knowledge of context, exposure to the speaker's L1, and ESL teacher status are most important for understanding low-comprehensibility speakers but become steadily less important as comprehensibility level increases until the highest level, when any listener can understand the L2 speech regardless of their background or knowledge of context.

Taken together, these results suggest that experienced listeners draw on several factors when judging L2 comprehensibility. These factors include aspects of grammar, vocabulary, and fluency in L2 speech and, at least for some listeners, word stress, discourse structure of the speaker's narratives, and the availability of context and familiarity with the speaker's L1.

Phase Three: Generating a L2 Comprehensibility Scale

The goal of the final phase was to identify the L2 speech measures that distinguish between three levels of L2 comprehensibility, and to distill verbal descriptions based on

those measures in an empirical L2 comprehensibility scale.

Selecting Measures

Of the 19 speech measures analyzed here, 18 significantly correlated with mean L2 comprehensibility ratings (see Table 1). Clearly, it was not feasible to include all these criteria in the rating descriptors, as it would not be practical for raters (or classroom teachers) to consult a long list of features when assessing L2 comprehensibility. Thus, the aim was to include only those measures that were both most closely related to the scores listeners assign and also most salient to them. Therefore, two criteria were applied to reduce the number of measures to be included as descriptors in the scale. The first (quantitative) criterion was that the correlations between the novice raters' L2 comprehensibility ratings and the speech measures from the quantitative phase needed to exceed .70, since this value conventionally designates strong associations. The second (qualitative) criterion was that the selected measures needed to have some conceptual link with a coded category in the ESL teachers' written reports from the qualitative phase. If a given measure was absent from the reports, then this would suggest that the teachers either did not ascribe much importance to it or were unable/unwilling to articulate it (Green, 1998).[4]

On the basis of the first criterion, five measures were retained: (1) type frequency, (2) word stress error ratio, (3) rhythm ratio, (4) MLR, and (5) story breadth (see Table 1). Token frequency was discarded due to its extremely high correlation with type frequency (r = .96), which suggests that the two frequency counts were not independent. Based on the second criterion, rhythm was excluded because the teachers did not comment on this variable. The remaining four measures did feature in the teacher reports. There was some correspondence between type frequency and both the coded categories of "vocabulary"

and "inadequate words produced;" between word stress error ratio and "word stress;" between MLR and "fluency;" and, finally, between story breadth and "storytelling elements and cohesion" (cf. Tables 1 and 2).

One intriguing finding was that grammatical error ratio, which was the first variable below the cutoff in the quantitative analysis ($r = -.63$), showed the clearest pattern in the teacher reports. All three teachers commented on grammar: it came up more frequently than any other coded category, and the overall pattern was clear in that the lower the comprehensibility level, the more grammar comments were made. Because grammar was important from the perspective of all three teachers, this measure was retained. Thus, five measures were finalized for inclusion in the rating scale: (1) type frequency, (2) word stress errors, (3) MLR, (4) story breadth, and (5) grammatical errors. The intercorrelations between these measures are shown in Table 3.

Table 3. *Pearson Correlation Coefficients Between the Speech Measures Selected for Inclusion in the Rating Scale*

	Type frequency	Word stress	MLR	Story breadth	Grammar
Type frequency	1				
Word stress	–.55**	1			
MLR	.88**	–.52**	1		
Story breadth	.74**	–.54**	.67**	1	
Grammar	–.45**	.45**	–.47**	–.36*	1

Note. *$p < .05$, **$p < .01$, two-tailed.

Distinguishing Between L2 Comprehensibility Levels

To examine whether the retained speech measures could distinguish between L2

speakers rated at low, intermediate, and high comprehensibility levels, five separate univariate analyses of variance (ANOVAs) were conducted, with comprehensibility level (low, intermediate, high) as the grouping factor and each of the retained speech measures as the dependent variable (Bonferroni corrected $\alpha = .01$). Table 4 shows that the means for all variables increased as L2 comprehensibility level increased, with the exception of the word stress and grammatical error measures, where error rates decreased as comprehensibility level increased ($p < .0001$). ANOVA statistics (also shown in Table 4) indicate that a medium to strong effect size was yielded for all measures ($\eta^2 = .34–.57$). The MLR and grammatical error results should be interpreted with caution, however, due to a violation of the assumption of homogeneity of variance.

The data were then submitted to Tukey HSD post-hoc tests to determine which of the three comprehensibility levels were different from one another for each L2 speech measure ($\alpha = .05$). Word stress, the measure with the largest effect size, distinguished between all three groups of L2 speakers. Significant differences between two of the three groups were found for the remaining four measures. Type frequency and MLR significantly distinguished between low- and intermediate-level speakers, whereas grammar errors and story breadth significantly distinguished between the high and intermediate groups. This suggests that a certain threshold of fluency and lexical diversity may be a useful criterion for distinguishing speakers at the low end of the comprehensibility continuum. Few grammatical errors and a large number of storytelling elements, on the other hand, may describe speakers at the high end of the continuum. The overall level distinctions based on these pairwise comparisons are summarized in Table 4.

Table 4. *Mean Scores (Standard Deviations) for the Selected Speech Measures Grouped by L2 Speaker Comprehensibility Level and Results of One-Way ANOVAs*

Speech measure	Speaker comprehensibility level			ANOVA results	
	Low	Intermediate	High	$F(2, 37)$	Effect size
Word stress error ratio	.57 (.17)	.30 (.22)	.10 (.32)	24.01	.57
Type frequency	.72 (.25)	1.18 (.20)	1.29 (.32)	17.31	.48
MLR	4.63 (1.61)	9.23 (3.17)	11.61 (5.22)	15.16	.45
Story breadth	.12 (.06)	.16 (.05)	.24 (.07)	11.28	.38
Grammatical error ratio	.13 (.08)	.08 (.05)	.04 (.03)	9.43	.34

Note. Speaker comprehensibility levels are based on the 60 novice raters' scalar judgments of L2 comprehensibility. Effect sizes are eta squared. All *F*-values are significant at $p < .0001$.

Developing L2 Comprehensibility Descriptors

The significant level distinctions (shown in Table 5) formed the basis for elaborating detailed descriptors for each of the three levels of L2 comprehensibility. The writing of the descriptors was informed by cumulative evidence from previous phases of the study. For example, the high intercorrelation between type frequency and MLR shown in Table 3 ($r = .88$) led to describing these aspects of speech in tandem. Teachers' comments from the written reports were also taken into account. For instance, T2's notion of comprehensibility based on listeners' background and knowledge of context were directly incorporated into the scale, although these observations are tentative due to the lack of quantitative evidence. The resulting descriptors perpetuate the illusion that all language features develop progressively (Fulcher et al., 2010; North, 2000) and do not make explicit the relative weightings of the linguistic influences on raters' comprehensibility judgments (Barkaoui, 2010). These are necessary considerations when

the scale is refined and validated in future research. A fully elaborated description of the scale appears in the appendix. Table 6 shows the simplified and therefore more user-friendly version of the scale.

Table 5. *Speech Measures that Distinguish Between Three Levels of L2 Comprehensibility*

Comprehensibility level	Speech measures		
High	Word stress	Type frequency MLR	Story breadth Grammar
Intermediate	Word stress		Story breadth Grammar
Low	Word stress	Type frequency MLR	

Note. The dotted lines separate the speech measures that significantly distinguish between the L2 speakers' comprehensibility levels (according to Tukey HSD post-hoc tests).

Table 6. *L2 Comprehensibility Scale with Simplified Descriptors*

Comprehensibility	The L2 speaker
High	• produces fluent stretches of speech; generally only pauses or hesitates at the end of the clause • provides sufficient vocabulary to propel the story forward or set the scene; lexical errors, if present, are not distracting • assigns word stress correctly in most instances • produces few grammatical errors that are unlikely to detract from the overall message
Intermediate	• produces some fluent stretches of speech; occasionally pauses or hesitates in the middle of the clause • experiences occasional lapses in vocabulary, although may roughly convey what the story is about; lexical errors are prevalent • is inconsistent in word stress placement • produces some grammatical errors that may detract from the overall message
Low	• produces dysfluent stretches of speech; frequently pauses or hesitates between lexical items • experiences frequent lapses in vocabulary that make the story indecipherable; high proportion of lexical errors, including L1 lexical influences • misplaces word stress in most instances • produces frequent grammatical errors that are likely to detract from the overall message

Discussion

Comprehensibility Level Distinctions

The goal of the present study was to investigate the factors that discriminate between different levels of L2 comprehensibility, for the purpose of developing an empirically-based rating scale. Overall, comprehensibility, which to date has been mostly investigated in L2 pronunciation and fluency studies (e.g., Kennedy & Trofimovich, 2008; Derwing et al., 2004), appears to be broader in its scope than previously thought.

Story breadth, for example, which was strongly correlated with L2 comprehensibility ratings, relates to both discourse organization (e.g., the use of cohesive devices) and pragmatic skills (e.g., identification of the story's referent; see de Villiers, 2004). This measure is likely specific to the particular picture-based narrative task used in the present study, and may not be relevant for word- or sentence-level tasks (see Kennedy, 2009). Nonetheless, the finding that story breadth is associated with L2 comprehensibility suggests that a wide range of measures feeds into listeners' comprehensibility ratings.

Overall, five speech measures were represented in the comprehensibility scale, with coverage from all four conceptual categories of phonology, fluency, linguistic resources, and discourse. Two measures (type frequency, MLR) distinguished between learners at the low end of the comprehensibility continuum. It may be that a certain threshold of lexical richness and fluency is required for learners to receive mid- to upper-range comprehensibility scores. Learners confined to the lowest comprehensibility level may not be able to retrieve lexical items efficiently, which impedes their ability to produce fluent stretches of speech and to convey a story in a short timeframe. At the opposite end of the spectrum, the higher-order skills of grammar and discourse organization (grammatical accuracy, story breadth) distinguished between only high-level learners. Evidence from the written reports showed that grammar errors were less likely to distract listeners from attending to the message as comprehensibility level increased. Likewise, T1's comments about the storyline were more frequent at the highest level, indicating that discourse organization mattered most for high-comprehensibility learners.

The word stress measure distinguished most strongly between the three comprehensibility levels in this study. Word stress is not contrastive (non-phonemic) in French, and Francophone learners often have difficulty perceiving L2 stress contrasts

(e.g., Peperkamp & Dupoux, 2002), which may also lead to production difficulties. L1 effects, therefore, likely come into play with word stress and possibly also rhythm, as suggested by a significant association between these measures ($r = -.62$). Both capture the speaker's ability to emphasize stressed syllables and reduce unstressed ones. It is possible that word stress, as a measure distinguishing most clearly between Francophone learners at different levels of L2 comprehensibility, is specific to these participants. However, judging from the sheer number of learners from other language backgrounds for whom English word stress (and rhythm) generally pose a problem (e.g., Spanish, Polish), stress patterns of English could be a much more global feature in distinguishing between different L2 comprehensibility levels (see Swan & Smith, 2001).

The robustness of the relationship between comprehensibility and word stress in this study throws into question the lack of emphasis on this and other suprasegmental aspects of L2 speech in Jenkins' (2000, 2002) Lingua Franca Core. This "pronunciation syllabus," based on observational research on communication breakdowns between nonnative dyads, comprises a list of instructional targets to be emphasized in a new, international variety of English. Notably, the native listeners in the present study are different from the nonnative interlocutors in Jenkins' work, and the speaking task here is nonreciprocal. However, previous research suggests that displaced stress patterns interfere with understanding for both native and nonnative listeners (Field, 2005), which argues for the importance of word stress for L2 comprehensibility.

In the present study, there was a link between word stress and story breadth (see Table 3), such that more word stress errors were associated with fewer propositions produced. This association likely arises because word stress (and rhythm) issues create a "bottleneck" at the phonological encoding and articulation stage of speech production

(Levelt, 1989; Segalowitz, 2010). The resulting slowdown is captured in temporal measures such as MLR and adversely affects comprehensibility. Learners may not have trouble with the message itself; they know what story elements need to be said (indeed, the images tell a clear story). Rather, learners struggle with "packaging" these story elements into appropriate words, and learners' inability to produce appropriate stress patterns may be a contributing factor. Clearly, the relationship between these variables and comprehensibility would benefit from further empirical work.

Segmental errors did not feature prominently in listeners' comprehensibility ratings in this study. One possibility is that the speakers had few segmental issues that contributed to difficulties in listener understanding. Alternatively, the measures examined here may not have been sensitive enough to capture segment-related errors leading to comprehension difficulties. Munro and Derwing (2006), for example, showed that errors involving consonants with a high functional load, that is, those that distinguish many lexical items (e.g., /l/ vs. /r/ in English), have a strong effect on comprehensibility whereas low functional load errors (e.g., /θ/ vs. /ð/) have only a minimal effect. Therefore, future research could take a more nuanced approach to examining the impact of segments on L2 comprehensibility. This could be achieved by focusing only on high functional load errors or by "zooming in" on listeners' reports of communication breakdowns to probe whether segmental errors may have played a part (see Zielinski, 2008).

Raters' Perspectives in the Scale Development Process

Pollitt and Murray (1996) argue that an investigation into raters' perceptions of proficiency should be the starting point of proficiency scale development. The same could be argued for comprehensibility. The approach adopted in the present study was to

initially probe raters' understanding of the construct without providing external guidance or criteria. The next step was to draw on these perspectives to ensure that the developed scale not only reflected the most statistically robust measures in relation to comprehensibility, but also included listeners' views on what they attended to when rating comprehensibility.

The ESL teacher raters showed individual differences in their interpretations of comprehensibility and in the criteria they found most salient while scoring. T1 commented on storytelling elements and cohesion. T2 indicated that L1 familiarity, L2 teacher status, and contextual support were necessary for the listener to understand the message. Finally, T3 drafted a formulaic descriptor listing the aspects of speech that had compromised her understanding of words, and was alone in citing word stress and intonation as being problematic. Clearly, defining comprehensibility in terms of ease of understanding leaves much leeway for interpretation and gives raters considerable freedom in choosing the speech characteristics to attend to when assigning scores. Therefore, raters in both research and assessment contexts would benefit from more direction on how ease of understanding is to be interpreted.

In the present study, it was not feasible to accommodate both a word-level and a story-based definition of comprehensibility in a single rating scale, as these entail different units of measurement (understanding words vs. longer texts), reflect different speaking tasks (reading out sentences vs. narrating a story), and likely involve a different set of measures contributing to comprehensibility. Ultimately, the inclusion of story breadth (which refers to the number of different propositions or story elements produced) as a criterion in the rating scale necessitated a story-based interpretation of comprehensibility. Some raters within the larger group of novice raters may have adopted

a word-level definition of comprehensibility, as did T3. However, when the novice raters' judgments were pooled and the teacher raters' comments were analyzed, the overall consensus was that the speaker's ability to convey the events of the story was a factor to consider when assigning comprehensibility scores.

Implications and Future Research

Although it is widely agreed that the goal of L2 pronunciation instruction should be to help learners be understandable to their interlocutors, classroom teachers have received little guidance on the pronunciation features to prioritize in instruction (Derwing & Munro, 2009). While not directly intended to inform instructional targets, the developed scale does point to the aspects of speech that listeners attend to when judging L2 comprehensibility. For example, the scale descriptors suggest that teaching learners to become more comprehensible involves not only specific pronunciation features (e.g., word stress) but also other language skills (e.g., vocabulary, discourse organization), since these are interrelated and are linked to listeners' perceptions of comprehensibility. Overall, the scale is intended as a formative assessment tool and could help interweave classroom-based oral assessment, including assessment of pronunciation, with L2 teaching and learning (Colby-Kelly & Turner, 2007).

Because the teacher raters who participated in this study instruct students from essentially the same population of learners as the Francophone speakers who provided speech samples, factoring teachers' decisions into the rating scale enhances its ecological validity. What is unclear is whether the linguistic aspects included in the scale are specific to Francophone learners or can be generalized to learners from other L1 backgrounds. It is therefore important to validate the scale for different groups of learners (e.g., from different L1 backgrounds, proficiency levels) and for different task-types (e.g.,

monologic, dialogic). It is also important to seek additional input from different rater groups with whom L2 speakers are likely to interact (e.g., ESL teachers, prospective coworkers). Because comprehensibility is frequently invoked in high-stakes assessment instruments (e.g., TOEFL) and is important for successful communication, there is a great need to develop a better understanding of comprehensibility. Investigations of the effects of systematic sources of variance on rating outcomes could reveal which of the criteria included in the scale are "stable" and generalize to other contexts, and which tend to be "local" and fluctuate across contexts (Chalhoub-Deville, 1996).

Concluding Remarks

Current theories of communicative ability may fall short of providing enough detail to guide test developers on how to model L2 pronunciation in oral proficiency scales, but some research findings have clear implications for rating scale development. One shortcoming of existing L2 oral proficiency scales, for example, is that comprehensibility and accentedness are often conflated in the descriptors, even though previous research has shown them to be partially independent dimensions (Derwing & Munro, 2009). Levis (2005) points out that the principle that L2 learners should simply strive to be understandable to their interlocutors is fundamentally incompatible with the idea that L2 learners should aim to acquire a nativelike accent, eradicating all traces of their L1. Rating scales need to reflect this reality. A research priority, therefore, should be to isolate the aspects of L2 speech that impede comprehensibility from those that, while noticeable or irritating, do not detract from listeners' understanding of the message (Munro, 2008). A recent study by Isaacs & Trofimovich (2010), for example, suggests that university-trained musicians are more sensitive to certain aspects of L2 speech than non-musicians, with the consequence that musicians tend to more clearly differentiate

between accentedness and comprehensibility than listeners who are less musically sensitive. Enlisting the perspectives of musically sensitive raters could, therefore, help tease apart these overlapping constructs. Once this has been accomplished, comprehensibility can be described in rating scales with greater precision, and reference to accent or nativelikeness can be left aside.

Acknowledgments

This research was supported by Social Sciences and Humanities Research Council of Canada (SSHRC) and Sir James Lougheed Award of Distinction doctoral fellowships awarded to the first author and by SSHRC and Fond québecois de la recherche sur la société et la culture (FQRSC) grants awarded to the second author. We are extremely grateful to Ron Thomson, Carolyn Turner, and Sarita Kennedy, whose input and insight contributed substantially to the content of this paper, and to Tracey Derwing and Murray Munro for the use of their speech elicitation materials. We also thank our research assistants Hyojin Song, Yvette Relkoff, Cassandre McLean Ikauno, Margaret Levey, Kathryn MacFadden-Willard, Fabrizio Stendardo, Joseph Hartfeil, and all our participants.

References

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.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*, 1–42.

Boersma, P., & Weenink, D. (2010). *Praat: Doing phonetics by computer* [computer program] (version 5.1.29). Available: www.praat.org

Breiner-Sanders, K., Lowe, P., Miles, J., & Swender, E. (2000). ACTFL proficiency guidelines - Speaking revised 1999. *Foreign Language Annals*, *33*, 13–18.

British Council, IDP: IELTS Australia, & UCLES, (n.d.). *IELTS Speaking band descriptors (public version)*. Retrieved June 24, 2010, from www.ielts.org/PDF/UOBDs_SpeakingFinal.pdf

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks*. TOEFL Monograph 29. Princeton, NJ: Educational Testing Service.

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.

Chalhoub-Deville, M. (1996). Performance assessment and the components of the oral construct across different tests and rater groups. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. 55–73). Cambridge: Cambridge University Press.

Colby-Kelly, C., & Turner, C. E. (2007). AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *Canadian Modern Language Review*, *64*, 9–37.

Cobb, T. (2000). The compleat lexical tutor [website]. Available: http://www.lextutor.ca

Cooper, W. E., & Sorensen, J. (1981). *Fundamental frequency in sentence production*. New York: Springer.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Couper, G. (2006). The short and long-term effects of pronunciation instruction. *Prospect*, *21*, 46–66.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31–51.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*, 1–16.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*, 1–15.

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

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., & Rossiter, M. J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning*, *13*, 1–17.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*, 665–679.

Deterding, D. (2001). The measurement of rhythm: A comparison of Singapore and British English. *Journal of Phonetics*, *29*, 217–230.

de Villiers, P. A. (2004). Assessing pragmatic skills in elicited production. *Seminars in Speech and Language*, *25*, 57–72.

Elliott, A. R. (1997). On the teaching and acquisition of pronunciation within a communicative approach. *Hispania*, *80*, 95–108.

ETS. (2005). *TOEFL iBT tips: How to prepare for the next generation TOEFL test and communicate with confidence*. Princeton, NJ: Author.

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

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

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, *18*, 299–323.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*, 208–238.

Fulcher, G. (2008). Criteria for evaluating language quality. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., pp. 157–176). New York: Springer.

Gaies, S. J. (1980). T-Unit analysis in second language research: Applications, problems and limitations. *TESOL Quarterly*, *14*, 53–60.

Gor, K., & Vatz, K. (2009). Less commonly taught languages: Issues in learning and teaching. In M. H. Long & C. J. Doughty (Eds.), *The handbook of second language teaching* (pp. 234–249). Oxford: Wiley-Blackwell.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press.

Harding, L. (in press). Pronunciation assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ: Wiley-Blackwell.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*, 24–49.

Isaacs, T., & Thomson, R. I. (2009, March). *Judgments of L2 comprehensibility, accentedness, and fluency: The listeners' perspective*. Paper presented at the Language Testing Research Colloquium, Denver, CO.

Isaacs, T. & Thomson, R. I. (2010). Revisiting research conventions: Rater experience, rating scale length, and judgments of L2 speech. Manuscript under revision.

Isaacs, T., & Trofimovich, P. (2010). Falling on sensitive ears? The influence of musical ability on extreme raters' judgments of L2 pronunciation. *TESOL Quarterly, 44*, 375–386.

Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of L2 speech. *Applied Psycholinguistics, 32*, 113–140.

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

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

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

Kelly, L. G. (1969). *25 centuries of language teaching: An inquiry into the science, art, and development of language teaching methodology, 500 B.C. –1969*. Rowley, MA: Newbury House.

Kennedy, S. (2009). L2 proficiency: Measuring the intelligibility of words and extended speech. In A. G. Benati (Ed.), *Issues in second language proficiency* (pp. 132–146). London: Continuum.

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

Krashen, S. (1981). *Second language acquisition and second language learning*. Oxford: Pergamon Press.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322.

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

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

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

Levis, J. M. (2006). Pronunciation and the assessment of spoken language. In R. Hughes (Ed.), *Spoken English, TESOL and applied linguistics: Challenges for theory and practice* (pp. 245–270). New York: Palgrave Macmillan.

Lippi-Green, R. (1997). *English with an accent: Language, ideology and discrimination in the United States*. London: Routledge.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Martin, J. R., & Rose, D. (2003). *Working with discourse: Meaning beyond the clause*. London: Continuum.

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

Morley, J. (1994). A multidimensional curriculum design for speech-pronunciation instruction. In J. Morley (Ed.), *Pronunciation pedagogy and theory* (pp. 64–91). Alexandria, VA: TESOL.

Munro, M. J. (2003). A primer on accent discrimination in the Canadian context. *TESL*

Canada Journal, *20*, 38–51.

Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. Hansen Edwards & M. Zampini (Eds.), *Phonology and second language acquisition* (pp. 193–218). Amsterdam: John Benjamins.

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

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

North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang.

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

Peperkamp, S., & Dupoux, E. (2002). A typological study of stress 'deafness'. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology 7* (pp. 203–236). Berlin: Mouton de Gruyter.

Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly*, *35*, 233–255.

Pollitt, A., & Murray, N. L. (1996). What raters *really* pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment. Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. 74–91). Cambridge: Cambridge University Press.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of non-native speaker conversations. *Discourse Processes*, *14*, 423–441.

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

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*, 93–120.

Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In R. Freedle (Ed.), *New directions in discursive processing* (pp. 53-120). Norwood, NJ: Ablex.

Swan, M. (1997). The influence of the mother tongue on second language vocabulary acquisition and use. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 156–180). Cambridge: Cambridge University Press.

Swan, M., & Smith, B. (Eds.). (2001). *Learner English: A teacher's guide to interference*. Cambridge: Cambridge University Press.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks, CA: Sage.

Terrell, T. D. (1989). Teaching Spanish pronunciation in a communicative approach. In P. C. Bjarkman & R. M. Hammond (Eds.), *American Spanish pronunciation: Theoretical and applied perspectives* (pp. 196–214). Washington, DC: Georgetown University Press.

Trofimovich, P., & Gatbonton, E. (2006). Repetition and focus on form in processing L2 Spanish words: Implications for pronunciation instruction. *Modern Language Journal, 90*, 519–535.

Trofimovich, P., Gatbonton, E., & Segalowitz, N. (2007). A dynamic look at L2 phonological learning: Seeking processing explanations for implicational phenomena. *Studies in Second Language Acquisition, 29*, 407–448.

UCLES. (2008). *Certificate of Proficiency in English: Handbook for teachers*. Cambridge: University of Cambridge ESOL Examinations.

Upshur, J. A., & Turner, C. E. (1999). Systemic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing*, *16*, 82–111.

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

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

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, *36*, 69–84.

Endnotes

1. Derwing and Rossiter (2003) also examined a variety of factors (including morphosyntactic and lexical semantic measures and segmental, prosodic, and fluency measures) in relation to comprehensibility. However, they only reported findings on segmental errors and pauses. Thus, no conclusions can be drawn from that study about the contribution of morphosyntactic and semantic factors to L2 comprehensibility.

2. Measures of grammatical accuracy and complexity based on t-units were not examined here because the L2 speakers produced simple clause structures, with no instances of subordination or embedded clauses in the entire dataset. Therefore, a measure based on t-units would not have discriminated effectively among the L2 speakers in this study (see Gaies, 1980).

3. Following Iwashita et al. (2008), type token-ratio was not examined due to the short length of the speech samples, since the number of tokens produced could be affected by speaking rate but may not be commensurate with the number of types produced, a measure of lexical density.

4. Of course, it was possible that some potentially important measures were not captured in the coding scheme and frequency counts; barring other evidence, however, measures that did not surface in the teachers' reports were not considered good candidates for inclusion in the rating scale.

Appendix:L2 Comprehensibility Scale with Elaborated Descriptors

Level	Elaborated comprehensibility descriptors
High	• [**Most listeners** are likely **to understand** the speech regardless of their familiarity with the context or exposure to the speaker's L1.] • The speaker's **access** to a **wide range** of **vocabulary** facilitates the production of **fluent stretches** of speech to set the scene of the story or propel the plot forward. Any shortcomings (e.g., omission of minor details or odd lexical choice) do not detract from the listener's ability to **follow** the **story**. • In **most instances**, the speaker accurately **emphasizes** the **strong syllable** in multisyllabic words. The speaker's ability to articulate the words and reduce syllables may contribute to listener impressions of **fluid speech**. • The speech contains **few grammatical errors**. Any such errors, if present, **do not detract** from the listener's **understanding**.
Intermediate	• [Occasional lapses in listener **understanding** are likely to be **enhanced** by **context** clues or **exposure** to the speaker's **L1**.] • The speaker's **lexical range** is **broad enough** to ensure the production of an **adequate number of words** (a ratable sample). However, **lexical choice** may be **imprecise or inappropriate**, and the speaker may **not** get **far** in conveying the events of the **story** or describing the scene. Listener **effort** is required to bypass these difficulties and **extract meaning from** the **story**. • The speaker is **inconsistent** in accurately **emphasizing** the **strong syllable** in multisyllabic words. Some difficulties articulating the words and reducing syllables may contribute to listener impressions of a "**lack of fluidity**." • The speech is interspersed with **occasional grammatical errors**, some of which **may detract** from the listener's **understanding**.
Low	• [**Most** listeners are likely to **have difficulty understanding** regardless of their exposure to the speaker's L1. Context is necessary to foster understanding.] • The speaker's **lexical range** is **limited**, and **frequent pauses** or hesitations **interfere with** the production of **fluent stretches** of speech. In some cases, there may be **L1 lexical intrusions** due to **lapses** in **vocabulary**, and the production of an **insufficient number of words** may make the speech sample difficult to rate. These factors **impede** the listener's ability to understand a **coherent storyline**. • In most instances, the speaker **does not emphasize** the **strong syllable** in multisyllabic words. Frequent difficulties articulating the words and reducing syllables may contribute to listener impressions of "**choppy speech**." • The speech is **replete** with **grammatical errors**, some which **may detract from** the listener's **understanding**.

Note. Square brackets are used for observations from the teacher reports that could not be verified using a different data source. Key items are bolded.

Chapter 5 – Final General Discussion

Interfaces between L2 Pronunciation and Assessment

The papers in this thesis have brought to the fore several issues in L2 pronunciation and assessment. If indeed "pronunciation continues to be the EFL/ESL orphan" (Gilbert, 2010, p. 3), then to say that the area of pronunciation assessment has been vastly underresearched would be an understatement. The challenge in trying to make papers in L2 pronunciation assessment relevant to a broad audience of applied linguists, who may not have particular interest in either pronunciation or assessment, is compounded by the fact that these concepts carry considerable "baggage." To elaborate, reference to "pronunciation" may evoke thoughts of "tedious mechanical activity" involving decontextualized drills (Yule, 1990, p. 111), while, "assessment," when interpreted in its narrower sense of "testing," may evoke thoughts of "a formal, test-room setting" involving test-taker response to discrete-point items (Coniam & Falvey, 2007, p. 460).

Although very little research has been conducted in the area of L2 pronunciation assessment since the time of Lado (1961), this thesis has shown there is much more to L2 pronunciation assessment than simply the discrete-point testing of segments of the sort that Lado proposed. Thus, part of the research agenda pursued in this thesis has been to reinvigorate the conversation on L2 pronunciation assessment within the academic community. The approach taken here has been (1) to identify research problems that need to be investigated with some urgency, (2) to address these problems using the methodological tools necessary to do so (i.e., following a pragmatic approach to research; see Teddlie & Tashakkori, 2009), (3) to "package" the findings into individual articles in a way that is relevant and accessible with particular journal readerships in mind, and

finally, (4) to strive to disseminate these papers in highly-cited journals that are likely to engage applied linguists with wide-ranging interests and, thereby, raise the profile of key issues in L2 pronunciation assessment.

Study 1, which was primarily written for a psycholinguistic audience, examines the cognitive variables of musical ability, phonological memory, and attention control, but not in order to explain variability in L2 learning gains or proficiency (e.g., O'Brien et al., 2006; Trofimovich et al., 2007a). Rather, the novel element of this paper is that these variables are examined as *rater* characteristics. The overall finding that these rater cognitive variables exerted little influence on raters' quantitative scoring decisions is reassuring, as this suggests that they are not a significant source of rater bias. *Study 2*, which was written primarily for language assessment professionals, who may be unfamiliar with L2 pronunciation research, brings together the L2 pronunciation literature with the assessment literature in building a case for the study. For example, Southwood and Flege's (1999) finding that at least 9-point accentedness scales are necessary to prevent a ceiling effect, which has influenced current research practice, directly contrasts with Alderson's (1991) contention that nine levels of pronunciation in a rating scale introduces unusable distinctions. Although these recommendations were made in reference to assessments for different purposes and with different stakes (i.e., low vs. high stakes, respectively), the number of levels that raters are reliably able to distinguish bears consideration in both contexts. Overall, the results suggest that for L2 assessment purposes, rating scales need to be more directive in helping raters understand the construct being measured and the most salient linguistic features at different levels of the scale (Pollitt & Murray, 1996). Clearly, numerical rating scales are limiting in this regard. This spurred the development of the L2 comprehensibility scale in *Study 3*, which took

raters' perceptions and scalar judgments into account in rating scale development. The 19 measures used to analyze the L2 speech data expanded the measures that Iwashita, Brown, and McNamara (2008) used to analyze L2 proficiency by including more objective and fine-grained segmental and suprasegmental measures and discourse-level measures. These measures are likely to be of interest to both SLA researchers and language testers who are interested in analyzing and quantifying linguistic features of L2 speech productions. Results suggested that comprehensibility cuts across a wide range of linguistic domains, including discourse and pragmatics, as was evidenced by the inclusion of the story breadth measure in the rating scale.

"What is the Construct?"(Bachman, 2007, p. 41)

Mirroring the papers included in this thesis, the issues raised in this closing chapter all relate to constructs, which are foundational to interpreting the scores derived from any L2 assessment in a construct-centered approach (Bachman & Palmer, 2010). There are several gaps in the present state of L2 pronunciation assessment research that limit the precision with which constructs can be defined, and the quality of validity evidence that can be gathered to support the proposed interpretations of the scores. First, existing theoretical frameworks do not adequately account for the role of pronunciation within the broader construct of communicative competence (Canale & Swain, 1980) or communicative language ability (Bachman, 1990; Bachman & Palmer, 1996). This absence of communicatively-oriented theory limits the substantive basis for articulating constructs definitions. As a result, several holistic constructs have been defined mostly based on the way they are operationalized (e.g., intelligibility entails listener transcriptions, whereas comprehensibility entails listener ratings; see Munro & Derwing,1999), a practice that Borsboom (2005) views as problematic due to the

inseparability of the construct from the measurement instrument. Another limitation that stems, in part, from this weak theoretical basis is the inconsistent or nebulous treatment of pronunciation in some L2 speaking scales (e.g., ACTFL) or its exclusion altogether (e.g., the CEFR Common Reference Scales). This implies that pronunciation is not an important component of L2 oral proficiency that needs to be consistently modeled in L2 oral proficiency scales, and makes it likely that pronunciation will act as a stealth factor during the rating process (Galaczi & Khalifa, 2009; Levis, 2006).

In this thesis, the approach to understanding more about the perceptual constructs of L2 comprehensibility, accentedness, and fluency was to consult the listener. This involved examining listeners' quantitative ratings in conjunction with their qualitative comments about the rating process and factors that influence their judgments, and then using these data to reflect back on our understanding of these constructs. In addition, in Study 3, the linguistic measures used to analyze the L2 speech samples were examined in relation to the rater data to determine which measures are most robust in making L2 comprehensibility level distinctions for rating scale development.

One theme that is pervasive in the papers that make up this thesis is that comprehensibility matters more than accentedness. This message comes across in each paper in a different way, and is discussed below with respect to the operationalization of constructs in rating scales, issues of reliability and validity, and methodological choices made in relation to the native speaker standard.

Comprehensibility and Accentedness: Not on an Equal Footing

As part of the renewed interest in L2 pronunciation, several researchers have advocated a shift in instructional focus away from acquiring a native-like accent (i.e., accent reduction) to the overall goal of intelligibility (Morley, 1991; Pennington &

Richards, 1986). In fact, the view expressed by an experienced rater in Study 2 that "I don't think an accent matters very much if you can understand the point the person's trying to make," is congruent with most L2 pronunciation researchers' views on the matter, including the authors of the papers that make up this thesis. It follows that intelligibility and comprehensibility, rather than accent, are "key to pronunciation assessment" (Levis, 2006, p. 252). Study 3 clarifies why "comprehensibility" is the appropriate term to use in reference to rating scales, rather than "intelligibility," in adhering to Derwing and Munro's (1997) definitional distinction (i.e., due to the way comprehensibility is operationalized). The relegation of accentedness to an inferior status relative to comprehensibility comes across in the papers in this thesis in the premise of the study (Study 3), expressed opinions of the raters (Study 2), and interpretation of the results (Study 1).

Although accentedness and comprehensibility stem from "contradictory principles" (Levis, 2005, p. 370), several rating scales conflate these dimensions in their descriptors (e.g., Cambridge ESOL Common Scale for Speaking). Scales that directly juxtapose incremental increases in comprehensibility with incremental decreases in accentedness lack empirical grounding (e.g., Speech Intelligibility/Communicability Index). In fact, as Study 3 shows, a wide range of linguistic factors appear to be relevant to comprehensibility level distinctions, including lexical and fluency measures at the low end of the scale (type frequency, MLR), grammatical and discourse-level measures at the high end of the scale (grammatical accuracy, story breadth), and word stress at all three levels of the scale. As noted in the discussion of Study 3, the word stress finding in particular may be L1-specific or applicable mostly to learners from syllable-timed languages. In addition, lexical, grammatical, and discourse-level measures can be

controlled in read-aloud tasks. Future research is needed to establish the generalizability of the criteria included in the scale to different groups of L1 speakers, raters, and speaking tasks and task types. What is clear from Study 3 without the need for future research, however, is that there is more to comprehensibility than accentedness alone, and that the criteria represented in a refined version of the scale are likely to be wide-ranging in terms of the linguistic domains represented.

The progression of papers in this thesis culminates in the development of the data-driven L2 comprehensibility scale in Study 3. This project could not have arisen had it not been for the precedents of Study 1 and, particularly, Study 2. In these latter studies, ratings of L2 accentedness and fluency were examined in conjunction with comprehensibility; no preferential treatment was given to any construct in the way the study was conceptualized or in the presentation of the constructs to raters, since they were all integral to the research questions. Study 2 will be discussed first. The purpose of this study was to revisit L2 pronunciation research conventions, some of which are evident in Study 1 (e.g., eliciting ratings of all three constructs on separate 9-point scales based on 20-30 second speech samples) by simulating these conventions, probing rater perceptions of the listening and rating process, and examining their scoring behavior as a function of rater experience and rating scale length. This led to reflections on current research practice.

One conclusion drawn from Study 2 was that numerical rating scales do not provide raters with sufficient guidance on how to interpret the constructs being measured and on what the different scale points "mean" in terms of performance quality. Although rating scales cannot be comprehensive in their inclusion of construct-relevant rating criteria (and the exclusion of a criterion does not *necessarily* mean that it is construct-

irrelevant), scales can at least set parameters on the scope of linguistic domains that raters should consider while rating (e.g., if the construct is limited to segmental and suprasegmental aspects of speech or also incorporates vocabulary, grammar, etc.).

As reported in Studies 1 and 2, Cronbach's alpha coefficients (or the numerically equivalent two-way mixed effects intraclass correlations, which are sometimes reported instead) tend to be high using the 9-point numerical rating scale (> .9), given a sample size of about 20 raters. However, these coefficients are likely inflated due to the homogeneous item format of the numerical rating scales (Tepper & Tepper, 1993). Furthermore, the Rasch category probability plots and rater comments suggest that raters do not apply rating scale levels consistently using this scoring method. This notwithstanding, even if a strong case is made for high interrater reliability by drawing on alternative ways of operationalizing it (Li, 2003), "reliability is a necessary but insufficient condition for validity" (Cohen, Manion, & Morrison, 2000, p. 105). That is, high interrater reliability coefficients do not necessarily mean that these ratings are valid. Validity issues, which are rarely addressed in current L2 pronunciation research, need to be brought to the fore, as has been done in this thesis.

In Study 2, the definition for fluency is the most directive of the three constructs examined. More specifically, a "narrow" definition of fluency based on temporal measures (Lennon, 1990)was provided along with examples of relevant features that could be taken into account in the ratings (e.g., filled and unfilled pauses, speech rate, etc.). Arguably, comprehensibility stands to benefit from empirical research of the sort that has been conducted on L2 fluency over the past several decades, including on task effects (e.g., Ejzenberg, 2000), empirical rating scale development (e.g., Fulcher, 1996), rater perceptions (e.g., Rossiter, 2009), and the relationship between listeners' global

ratings and temporal measures used to analyze the L2 speech samples (e.g., Derwing, et al., 2004).

In contrast to fluency, L2 comprehensibility stands out as the most ambiguous of the three constructs examined in Study 2 in terms of informing raters about what precisely is being measured. For instance, it is not clear whether the definition of "how easy the speaker is to understand" refers to listener understanding each individual word that the speaker enunciates, or understanding of the overall story or message. This broad definition does not lead to rater consensus, as is evident in the three teachers' divergent interpretations of comprehensibility in Study 3. In addition, raters' comments in Studies 2 and 3 suggest that the issue of whether or not they should assume prior knowledge of context (i.e., the picture prompt) when assigning scores needs to be specified. Similarly, the fact that some experienced raters put themselves "in the average person's shoes" when scoring rather than rating based on their own understanding as ESL professionals suggests that the issue of who the listener is (i.e., the target audience) also needs to be clarified. Notably, in response to one teacher's written comments in Study 3, contextual support and listener L2 teaching background were built into the elaborated L2 comprehensibility descriptors at each level of the scale (see square bracketed criteria in the appendix).

The inclusion of the story breadth measure in the scale based on the established criteria suggests that comprehensibility involves more than just the enunciation of individual words. Rather, discourse-level understanding (i.e., being able to follow the sequence of events in the story) should also be taken into account. Admittedly, this proposition-based measure would not have been possible had word-level, or sentence-level speech elicitation tasks been used or if the task consisted of a read-aloud (i.e.,

diagnostic passage) rather than an extemporaneous speaking task. So although task effects were not directly examined in this thesis, the influence of the storytelling task on L2 learners' productions is acknowledged through the inclusion of a proposition-based measure in the scale. In this limited sense, the rating scale is discernably task-based. Again, the issue of the contexts that the scale generalizes to, including tasks and task-types, needs to be systematically investigated in future research.

With respect to the relative importance of comprehensibility versus accentedness, some raters in Study 2 commented negatively on the use of the descriptor "not accented at all" at the high end of the accentedness scale because this appeared to be an unattainable standard. Other raters questioned the premise of rating accentedness in the first place. These comments led to a reflection on the inclusion of native speakers in the training samples to establish the upper bounds of the scale. The disparity between the native speakers in the sample ratings and the L2 speakers was particularly glaring due to the low proficiency of these learners, who were assessed at beginner levels of the Canadian Language Benchmarks.

Clearly, speech samples used for training purposes need to be rigorously benchmarked to rating scale levels. Rather than being used to carefully train raters, however, the training samples in Study 2 were primary used to familiarize raters with the procedure and reassure them that their global intuitions of the speech broadly conformed with ratings that a previous group of raters had assigned. It is difficult to conceive of benchmarking performance samples to the numerical rating scales used in Studies 1 and 2 for rater training purposes, due to the lack of verbal descriptors for scale points between rating scale anchors. Now that a preliminary descriptive L2 comprehensibility scale has

been developed in Study 3, establishing benchmark samples is essential if the instrument is eventually to be used for formative assessment purposes.

Raters' comments on accentedness and the native-speaker norm in Study 2 also prompted reflection on the methodological choice of using native speakers as "placeholders" in accordance with previous L2 speech research (e.g., Derwing, Munro, & Thomson, 2008; Derwing, Munro, & Wiebe, 1998). This refers to the practice of interspersing the L2 speech samples with performances of a few native speakers on the same speaking task (picture narrative) as a means of verifying that the scores raters assign correspond to the intended item number (speech sample). The assumption behind this practice is that the native speakers will be scored at the extreme high end of the scale, and this is invariably what takes place. Thus, the items featuring native speakers constitute checkpoints for ensuring that ratings and item numbers are not misaligned. Once rating-item correspondence has been verified, the native speaker ratings are discarded from the analysis, since they are not the subject of interest in the study and could skew results for the L2 speakers. Raters are not informed that these ratings do not count.

The major problem with this approach is that it encourages raters to compare the L2 speakers' performances with those of the native speakers and upholds an unrealistic standard for no strong substantive reason. In a recent study by Thomson and Isaacs (2010), an alternative approach was piloted that does not rely on native speaker placeholders but still addresses the practical need to ensure rating-item correspondence. When the L2 speech samples to be rated are burned onto CDs in randomized order, tracks after a fixed number of items are reserved for a "pleasant" sounding voice (not necessarily of a native-speaker) who says, "you should now be on item number __." While this approach is more artificial than the placeholder approach, it is also more

transparent and does not waste raters' time assigning ratings that will only be thrown out. More importantly, this approach does not encourage raters to use native speaker performances as a gold standard against which L2 speakers' performances are unfairly compared.

Study 2 provided raters with a platform to talk about their experience listening to and evaluating L2 speech. Some raters were forthcoming in their critique of accentedness and adherence to the native speaker standard, and these points were taken into account in the researchers' reflections on current conventions. Conversely, Study 1 is a purely quantitative study, and raters were not able to voice their opinions in this way. In this study, the subordinate role of accentedness relative to comprehensibility came through in the researchers' interpretation of the findings. These will be briefly summarized here to set the background for follow-up studies. Finally, the thesis will conclude with a discussion of future directions.

Study 1 is apparently the first study to examine whether individual differences in rater cognitive variables influence the scores they assign on measures of L2 comprehensibility, accentedness, and fluency. The two main findings were reassuring. The first finding was that individual differences in raters' phonological memory and attention control did not have a bearing on mean ratings they assigned on any of the perceptual measures. This null result suggests that these psycholinguistic variables do not pose a threat to the validity of the assessments, at least in terms of quantitative scoring outcomes. The second finding was that music majors assigned significantly lower mean ratings than non-music majors solely for accentedness, and particularly for heavily accented L2 speakers. This suggests that music majors tend to be more sensitive to certain

(as yet undefined) aspects of L2 speech that contribute to accentedness. No group differences were detected for comprehensibility or fluency ratings.

Of course, what really counts in communication is whether interlocutors are able to understand each other's speech, and not simply the presence of a detectable accent. As an experienced rater in Study 2 emphasized, "accent doesn't always make a difference in terms of [the speaker] being easy to understand" (i.e., does not necessarily interfere with understanding). In line with this, previous research has shown that an L2 learner who is difficult to understand is almost always judged as being heavily accented, whereas the reverse is not necessarily the case (Munro & Derwing, 1999). Thus, findings for accentedness should only be interpreted in reference to comprehensibility or intelligibility when extrapolating beyond the research context. Future research needs to investigate the relationship between raters' musical background and intelligibility with some urgency, hopefully to rule out the presence of a musical training effect, which would pose a serious threat to validity.

Study 1 opened up nearly as many new questions as it answered. In a follow-up study, Isaacs and Trofmovich (2010) examined a portion of the original data to extend the findings. The main analysis involved identifying the ten most severe and ten most lenient raters on each of the rated measures (L2 comprehensibility, accentness, and fluency) and conducting separate *t*-tests to examine group differences in rater performance on the Musical Aptitude Profile (MAP), which was the measure of musical ability used in this study. Phonological memory and attention control were not investigated due to nonsignificant effects in Study 1.

Surprisingly, no significant difference on MAP performance between extreme severe and extreme lenient raters was detected for accentedness as might have been

expected based on the findings from Study 1. Instead, a significant effect was detected solely for extreme comprehensibility raters. While differences were significant using the MAP composite score, when the subtests were examined separately, it became clear that the significant result was mostly due to differences on the MAP melody subtest. Raters who had indicated having the hardest time understanding the L2 speech were also the strongest performers on the melodic MAP subtest, as was evident in tightly clustered points on the scatterplot. This attention to melodic aspects of music and speech could presumably distract extremely severe raters from focusing on the L2 speaker's message, as reflected in their comprehensibility ratings. This would explain their perception of the speech being more difficult to understand relative to the perceptions of more moderate and lenient raters. At the other end of the spectrum, MAP melody performance for the 10 most lenient raters was much more diffuse, making it difficult to draw conclusions about musical ability for lenient raters.

Although findings between Study 1 and this follow-up study were somewhat contradictory in that the significant result involved accentedness in Study 1 and comprehensibility in Study 2, these differences can, in part, be accounted for by the different focus and level of detail examined in these studies. Study 1 examined differences in mean ratings assigned by 30 music majors and 30 non-music majors, whereas the follow-up study examined differences in musical ability between the 10 extremely severe and 10 extremely lenient raters. Despite these discrepancies, taken together, the overall trend is consistent with the narrative established in Study 1: raters with a high level of musical ability/experience tend to be more sensitive to certain aspects of pronunciation than raters with lower musical ability/less musical experience. There is some evidence in the follow-up study that musical raters' sensitivity to melodic

dimensions of music and speech accounted for these differences. Conversely, the failure of the tempo subtest (which relates to rhythm and timing) to detect differences between extreme rater groups could explain why L2 ratings of fluency appear to be unrelated to individual differences in musical ability/experience, as per the results of both studies.

If raters' musical ability plays a role in influencing their judgments of L2 comprehensibility, then this is potentially of far greater concern than effects on accentedness ratings, due to the importance of L2 comprehensibility as an assessment criterion in high-stakes tests and in governing the success of interactions among interlocutors in communicative settings. However, further research is clearly needed before making sweeping claims about the influence of musical ability on ratings of L2 comprehensibility. The follow-up study to Study 1, for example, involved only a small number of extreme raters. In addition, the slightly contradictory findings across the two studies suggest that the data were not extremely robust and need to be replicated. Finally, musical ability and experience variables need to be teased apart in future research, and a measure of the *quality* of an individual's musical experience might be interesting to implement, although it is not clear how this could be quantified.

Final Thought

Despite these caveats and challenges, the papers in this thesis adopt a unique approach to interfacing L2 pronunciation research with assessment research by investigating systematic sources of variance in the rating process, bringing to light issues of construct validity, and attempting to make these discussions relevant to applied linguists with wide-ranging research interests. It is through this sort of interdisciplinary research and its dissemination in different journals to different readerships that the field of pronunciation assessment may finally move beyond Lado (1961).

General References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 71–86). London: Macmillan.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–70). Ottawa, ON: University of Ottawa Press.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449–465.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*, 238–257.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*, 54–74.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 1111*, 1061–1071.

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

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17*, 5–9.

Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 112–140). Cambridge: Cambridge University Press.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*, 1–15.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1–25.

Brown, A., Iwashita, N., & McNamara, T. F. (2005). *An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks*. Monograph Series MS-29. Princeton, NJ: Educational Testing Service.

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.

Chalhoub-Deville, M. (1995). A contextualized approach to describing oral language proficiency. *Language Learning, 45*, 251–281.

Chalhoub-Deville, M. (1996). Performance assessment and the components of the oral construct across different tests and rater groups. In M. Milanovic & N. Saville

(Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. 55–73). Cambridge: Cambridge University Press.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*, 369–383.

Chapelle, C. A. (1998). Construct validation and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge: Cambridge University Press.

Chun, C. W. (2006). An analysis of a language test for employment: The authenticity of the PhonePass test. *Language Assessment Quarterly, 3*, 295–306.

Cohen, L., Manion, L., & Morrison, K. R. B. (2000). *Research methods in education* (5th ed.). London: Routledge Falmer.

Colby-Kelly, C., & Turner, C. E. (2007). AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *Canadian Modern Language Review, 64*(1), 9–37.

Coniam, D., & Falvey, P. (2007). High stakes testing and assessment: English language teacher benchmarking, Part 1. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 457–471). New York: Springer.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*, 31–51.

Cumming, A., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework*. TOEFL Monograph 22. Princeton, NJ: Educational Testing Service.

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

Derwing, T. M. (2008). Curriculum issues in teaching pronunciation to second language learners. In J. Hansen Edwards & M. Zampini (Eds.), *Phonology and second language acquisition* (pp. 347–369). Amsterdam: John Benjamins.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*, 1–16.

Derwing, T. M., & Munro, M. J. (2009a). Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *Canadian Modern Language Review, 66*, 181–202.

Derwing, T. M., & Munro, M. J. (2009b). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*, 1–15.

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

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

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

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., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*, 665–679.

Deville, C., & Chalhoub-Deville, M. (2006). Old and new thoughts on test score variability: Implications for reliability and validity. In M. Chalhoub-Deville, C. A. Chapelle & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple research perspectives* (pp. 9–26). Amsterdam: John Benjamins.

Dewaele, J.-M. (2009). Individual differences in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *The new handbook of second language acquisition* (2nd ed., pp. 623–646). Bingley, UK: Emerald Gold.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second langauge acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*, 125–144.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*, 155–185.

Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 287–313). Ann Arbor, MI: Univeristy of Michigan Press.

Ellis, R. (2004). Individual differences in second language learning. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 525–551). Malden, MA: Blackwell.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In Philological Society of London (Ed.), *Studies in Linguistic Analysis* (pp. 1–32). Oxford: Blackwell.

Fox, J., & Fraser, W. (2009). Test review: The Versant Spanish Test. *Language Testing, 26*, 313–322.

Franco, H., Bratt, H., Rossier, R., Gadde, V. R., Shriberg, E., Abrash, V., Precoda, K. (2010). EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*. Advance online publication. doi: 10.1177/0265532210364408

French, L. M., & O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics, 29*, 463–489.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*, 208–238.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson.

Fulcher, G. (2008). Criteria for evaluating language quality. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., pp. 157–176). New York: Springer.

Fulcher, G. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics, 29*, 3–20.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.

Fulcher, G., Davidson, F., & Kemp, J. (2010). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*. Advance online publication. doi: 10.1177/0265532209359514

Galaczi, E. D., & ffrench, A. (2007). Developing revised assessment scales for Main Suite and BEC Speaking tests. *Research Notes, 30*, 28–31.

Galaczi, E., & Khalifa, H. (2009). Cambridge ESOL's CEFR DVD of speaking performances: What's the story? *Research Notes, 37*, 23–29.

Gatbonton, E., & Segalowitz, N. (2005). Rethinking communicative language teaching: A focus on access to fluency. *Canadian Modern Language Review, 61*, 325–353.

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.

Gilbert, J. B. (2010). Pronunciation as orphan: What can be done? *Speak Out!, 43*, 3–7.

Goodwin, J., Brinton, D., & Celce-Murcia, M. (1994). Pronunciation assessment in the ESL/EFL curriculum. In J. Morley (Ed.), *Pronunciation pedagogy and theory: New views, new directions* (pp. 3–16). Alexandria, VA: TESOL.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*, 201–233.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning, 41*(3), 337–373.

Harding, L. (2008). Accent and academic listening assessment: A study of test-taker perceptions. *Melbourne Papers in Language Testing, 13*, 1–33.

Harding, L. (in press). Pronunciation assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ: Wiley-Blackwell.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *Journal of General Psychology, 123*, 207–215.

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English speaking graduate students. *Canadian Modern Language Review, 64*, 555–580.

Isaacs, T. (2009). Integrating form and meaning in L2 pronunciation instruction. *TESL Canada Journal*, *26*, 1–12.

Isaacs, T. (in press). Phonology: Mixed methods. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ: Wiley-Blackwell.

Isaacs, T., & Thomson, R. I. (2009, March). *Judgments of L2 comprehensibility, accentedness, and fluency: The listeners' perspective*. Paper presented at the Language Testing Research Colloquium, Denver, CO.

Isaacs, T., & Trofimovich, P. (2010). Falling on sensitive ears? The influence of musical ability on extreme raters' judgments of L2 pronunciation. *TESOL Quarterly, 44*, 375–386.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*, 24–49.

Jenkins, J. (1998). Which pronunciation norms and models for English as an International Language? *ELT Journal, 52*, 119–126.

Jenkins, J. (2006). The spread of EIL: A testing time for testers. *ELT Journal, 60*, 42–50.

Kane, M. (2010). Validity and fairness. *Language Testing, 27*, 177–182.

Kang, O. (2008). Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness. *Spaan Fellow Working Papers, 6*, 181–205.

Kennedy, S. (2009). L2 proficiency: Measuring the intelligibility of words and extended speech. In A. G. Benati (Ed.), *Issues in second language proficiency* (pp. 132–146). London: Continuum.

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

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*, 187–217.

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

Kunnan, A. J. (Ed.). (2008). *Towards a model of test evaluation: Using the test fairness and the test context frameworks*. Cambridge: Cambridge: UCLES/Cambridge University Press.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.

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

Leighton, J. P. (2008). Review of [Measuring the mind: Conceptual issues in contemporary psychometrics by D. Borsboom]. *Journal of Educational Measurement, 45*, 91–94.

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

Levis, J. M. (2006). Pronunciation and the assessment of spoken language. In R. Hughes (Ed.), *Spoken English, TESOL and applied linguistics: Challenges for theory and practice* (pp. 245–270). New York: Palgrave Macmillan.

Li, H. (2003). The resolution of some paradoxes related to reliability and validity. *Journal of Educational and Behavioral Statistics, 29*, 241–244.

Lindemann, S. (2006). What the other half gives: The interlocutor's role in nonnative speaker performance. In R. Hughes (Ed.), *Spoken English, TESOL and applied linguistics: Challenges for theory and practice* (pp. 23–49). Basingstoke, UK: Palgrave Macmillan.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education *Educational Researcher, 36*, 437–448.

Lord, G. (2008). Podcasting communities and second language pronunciation. *Foreign Language Annals, 41*, 364–379.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

MacDonald, S. (2002). Pronunciation: Views and practices of reluctant teachers. *Prospect, 17*, 3–18.

Markus, K. A. (1998). Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible? *Social Indicators Research, 45*, 7–34.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. F. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3*, 31–51.

McNamara, T. F., & Roever, C. (2006). *Language Testing: The social dimension*. Malden, MA: Blackwell.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.

Messick, S. (1990). *Validity of test interpretation and use*. Research Report 90-11. Princeton, NJ: Educational Testing Service.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher, 23*, 13–23.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241–256.

Messick, S. J. (1998a). Alternative modes of assessment, uniform standards of validity. In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternative to traditional testing for selection* (pp. 59–74). Mahwah, NJ: Lawrence Erlbaum.

Messick, S. (1998b). Test validity: A matter of consequence. *Social Indicators Research, 45*, 35–44.

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

Morley, J. (1994). A multidimensional curriculum design for speech-pronunciation instruction. In J. Morley (Ed.), *Pronunciation pedagogy and theory* (pp. 64–91). Alexandria, VA: TESOL.

Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher, 25*, 20–29.

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice, 22*, 13–25.

Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. Hansen Edwards & M. Zampini (Eds.), *Phonology and second language acquisition* (pp. 193–218). Amsterdam: John Benjamins.

Munro, M. J., & Derwing, T. M. (1994). Evaluations of foreign accent in extemporaneous and read material. *Language Testing, 11*, 254–266.

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

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning, 15*, 441–467.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang.

O'Brien, I., Segalowitz, N., Collentine, J., & Freed, B. (2006). Phonological memory and lexical, narrative, and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics, 26*, 377–402.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*, 169–192.

O'Loughlin, K. (2007). An investigation into the role of gender in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in*

speaking and writing assessment (pp. 63–97). Cambridge: Cambridge University Press.

O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition*. Cambridge: Cambridge University Press.

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

Pienemann, M. (2005). An introduction to Processability Theory. In M. Pienemann (Ed.), *Cross-linguistic aspects of Processability Theory* (pp. 1–60). Amsterdam: John Benjamins.

Pollitt, A., & Murray, N. L. (1996). What raters *really* pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment. Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. 74–91). Cambridge: Cambridge University Press.

Rajadurai, J. (2007). Intelligibility studies: A consideration of empirical and ideological issues. *World Englishes, 26*, 87–98.

Ranta, L. (2002). The role of learners' language analytic ability in the communicative classroom. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 159–180). Amsterdam: John Benjamins.

Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing, 18*, 429–462.

Robinson, P. (Ed.). (2002). *Individual differences and instructed language learning*. Amsterdam: John Benjamins.

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

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review, 65*, 395–412.

Saville, N. (2003). Continuity and innovation: Revising the Cambridge Proficiency in English Examination 1913-2002. In C. J. Weir & M. Milanovic (Eds.), *The process of test development and revision within UCLES EFL* (pp. 57–120). Cambridge: Cambridge University Press/UCLES.

Segalowitz, N. (1997). Individual differences in second language acquisition. In A. M. B. De Groot & J. Kroll (Eds.), *Tutorials in bilingualism* (pp. 85–112). Hillsdale, NJ: Lawrence Erlbaum.

Segalowitz, N., & Frenkiel-Fishman, S. (2005). Attention control and ability level in a complex cognitive skill: Attention-shifting and second language proficiency. *Memory and Cognition, 33*, 644–653.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*, 5–24.

Skehan, P. (1989). *Individual differences in second-language learning*. London: Edward Arnold.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency: Does musical ability matter? *Psychological Science, 17*, 675–681.

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

Sparks, R., & Ganschow, L. (2001). Aptitude for learning a foreign language. *Annual Review of Applied Linguistics, 21*, 90–111.

Szpyra-Kozłowska, J., Frankiewicz, J., Nowacka, M., & Stadnicka, L. (2005). Assessing assessment methods: On the reliability of pronunciation tests in EFL. Retrieved August 15, 2010, from www.phon.ucl.ac.uk/home/johnm/ptlc2005/pdf/ptlcp37.pdf

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks, CA: Sage.

Tepper, B. J., & Tepper, K. (1993). The effects of method variance within measures. *The Journal of Psychology: Interdisciplinary and Applied, 127*, 293–302.

Thomson, R. I. (in press). Accent reduction. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ: Wiley-Blackwell.

Thomson, R. I., & Isaacs, T. (2010, March). *The real me vs. artificial her: The effect of grammatical person, topic familiarity and task type on L2 oral performance*. Conference of the American Association of Applied Linguistics, Atlanta, GA.

Trofimovich, P., Ammar, A., & Gatbonton, E. (2007a). How effective are recasts? The role of attention, memory, and analytical ability. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 171–195). Oxford: Oxford University Press.

Trofimovich, P., Gatbonton, E., & Segalowitz, N. (2007b). A dynamic look at L2 phonological learning: Seeking processing explanations for implicational phenomena. *Studies in Second Language Acquisition, 29*, 407–448.

Turner, C. E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review, 56*, 555–584.

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

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*, 3–12.

Upshur, J. A., & Turner, C. E. (1999). Systemic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing, 16*, 82–111.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23*, 411–440.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Xi, X. (2008). Methods of test validation. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (Vol. 2, pp. 177–196). New York: Springer.

Yule, G. (1990). Review of [Teaching English pronunciation, by J. Kenworthy; Teaching pronunciation: Focus on English rhythm and intonation, by R. Wong; and Current perspectives on pronunciation: Practices anchored in theory, by J. Morley]. *System, 18*, 107–111.

Supplementary Materials to Thesis

Appendix A: Rater Background Questionnaire, Study 1

Name: _____ Gender: Male ___ Female ___

Date of birth: _____ Birthplace (City, Province/State): _____

Degree/ Major/ Year of study: _____
(e.g., BSc/ Chemistry/ 2nd year)

Is your hearing normal as far as you know? Yes: ___ No: ___

What do you consider to be your native language? French: ___ English: ___
Both: ___ Other: ___

If English is your native language, which dialect of English do you speak? _____
(e.g., Newfoundland, BC)

What language(s) have you been exposed to from birth?_____

What do you consider to be your second language? French: ___ English: ___
Both: ___ Other: ___

At what age did you start learning your second language? _____

What language do you speak at home now? _____

What is the native language of your mother? _____ Your father? _____

In what language did you attend school? Please circle the appropriate answer

- elementary school: French only English only French & English mix Other: _____

- high school: French only English only French & English mix Other: _____

- CEGEP (if applies): French only English only French & English mix Other: _____

Period of residence in Montreal: _____
(e.g., 2 years; whole life)

Period of residence in other French speaking environment(s) exceeding 1 month, if applicable:

Place: _____ Year: _____ Length: _____ Reason: _____
(e.g., France) (e.g., 2005) (e.g., 2 months) (e.g., bursary)

Place: _____ Year: _____ Length: _____ Reason: _____

Please rate your ability to speak, listen to, read and write **French** using the scales in the box below. *Note* that 1= extremely poor and 9= extremely fluent.

French: 1 = Extremely Poor 9 = Extremely Fluent

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

Please rate how well you speak, listen to, read and write **English** in the box below.

English: 1 = Extremely Poor 9 = Extremely Fluent

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

Please indicate the approximate percent of time that you use the following languages each week by circling the appropriate percent for each skill (speaking, listening, reading, writing).

French

Speaking	0%	10	20	30	40	50	60	70	80	90	100%
Listening to Media	0%	10	20	30	40	50	60	70	80	90	100%
Reading	0%	10	20	30	40	50	60	70	80	90	100%
Writing	0%	10	20	30	40	50	60	70	80	90	100%

English

Speaking	0%	10	20	30	40	50	60	70	80	90	100%
Listening to Media	0%	10	20	30	40	50	60	70	80	90	100%
Reading	0%	10	20	30	40	50	60	70	80	90	100%
Writing	0%	10	20	30	40	50	60	70	80	90	100%

Please indicate the approximate percentage of time that you **communicate in English** each week with individuals whose native language is not English (e.g., French speakers, Russian speakers, etc).

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

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

Do you have any teaching experience? Yes: ___ No: ___

If so, please describe the context:

Place: _____ Year: _____ Length: _____ Subject: _____
(e.g., Montreal) (e.g., 2005) (e.g., 1 year) (e.g., swimming)

Place: _____ Year: _____ Length: _____ Subject: _____

Have you had any pronunciation training or taken a phonology course? Yes: ___No: ___

If so, please describe the context:

Place: _____ Year: _____ Course: _____ Other info:_____

Please rate your own abilities in the following areas using the scales in the box below. *Note* that 1= extremely poor and 9 = extremely good.

1 = Extremely Poor **9 = Extremely Good**

Your ability to perceive/distinguish subtleties of music	Your ability to remember words spoken in language you do not know
1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9

1 = Extremely Poor **9 = Extremely Good**

Your ability to shift from one task your are working on to another (typing & speaking)	Your ability to imitate words spoken in a language you do not know
1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9

Have you had any formal musical training? Yes: ___ No: ___

If you answered NO, please stop here. If you are a music major, please proceed.

What is your primary instrument? _____
(e.g., piano)

How many years of formal training have you had in that instrument? _____
(e.g., 5 years)

What courses/lessons have you taken as part of training in that instrument?

Course/Lessons: _____ Year: _____ Length: _____
(e.g., youth choir) (e.g., 2005) (e.g., 5 months)

Course/Lessons: _____ Year: _____ Length: _____

Course/Lessons: _____ Year: _____ Length: _____

Course/Lessons: _____ Year: _____ Length: _____

If you have any accreditation in that instrument, please list here: _____
(e.g., Grade 8 Royal Conservatory piano)

Please list any other instruments that you play (if applicable), the number of years of performance practice that you have had, and any accreditation:

Instrument: _____ Years of practice: ___ Accreditation: _____

Instrument: _____ Years of practice: ___ Accreditation: _____

How many courses in ear training have you taken? _____

What is the most advanced ear training course that you completed?

_____ When? _____
(course title)

How many courses in theory/composition have you had? _____
(number)

What is the most advanced theory course that you have completed?

_____ When? _____
(course title)

Has most of your musical knowledge and training been in the Western classical tradition?

Yes: ___ No: ___

If not, what other musical traditions are you familiar with? Please explain.

_____
(e.g., Jazz musician, only classical training is in a first year theory course)

Appendix B: Background Questionnaire for Experienced Raters, Study 2

This purpose of this questionnaire is to gather information about your language and teaching background. Please answer as completely as you can by filling in the blanks or circling the best answer.

1. Birthplace (City, Province/State): _____

2. Age a. 20-29 b. 30-39 c. 40-49 d. over 50

3. Current Degree/ Major/ Year of study (if applicable):

_____
(e.g., MA/ Second language education/ 2nd year)

4. Last Degree you earned/ Major: _____

5. Is your hearing normal as far as you know? a. Yes b. No

Language Use and Background

6. What is your native language (from birth)? _____

7. What language did you do your schooling in? Please specify if "other."

- elementary school: a. English b. French c. Other: _____

- high school: a. English b. French c. Other: _____

- CEGEP: a. English b. French c. Other: _____

- university: a. English b. French c. Other: _____

8. Approximately what percent of the time do you speak English (as opposed to other languages) in your daily life?

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

9. Approximately what percent of the time do you listen to the English language media (as opposed to the media in other languages)?

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

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

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

11. What other languages do you know? _____

Which of these languages would you say that you are fluent in? _____

12. Have you ever lived in a non-English speaking country for more than 3 months?

a. Yes b. No

If so, please specify.

Country: _____ Time: _____ Reason:_____
(e.g., Italy) (e.g., 6 months) (e.g., university exchange)

Country: _____ Time: _____ Reason:_____

Country: _____ Time: _____ Reason:_____

13. How familiar are you with the spoken English of people from the following first language backgrounds? Note that 1 = extremely Unfamiliar; 9 = extremely familiar

Russian	**Mandarin**
1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9

1 = extremely unfamiliar; 9 = extremely familiar

14. Have you had significant exposure to people from either language group (Russian or Mandarin)?

a. Yes b. No If so, describe the context:_____

Teaching Experience and Training

15. How many years have you taught ESL? ___ years

16. Approximately how many hours of ESL do (did) you teach per week? ___ hours

17. What type of institution do (did) you teach ESL at? Circle more than one answer if applicable.

a. Private language institute

b. Language institute for college/university-bound students

c. College/University

d. Other (please specify): _____

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

a. Yes b. No

If so, please specify. Course: _____ Other info: _____

Course: _____ Other info: _____

19. Have you ever taken a language assessment course? a. Yes b. No

20. Have you ever been trained as a rater? a. Yes b. No

If so, please specify. Amount of training: _____ Skill(s) tested: _____
(e.g., 2 hours) (e.g., writing)

21. Have you ever used a rating scale to evaluate the students in your ESL classroom?

a. Yes b. No

If so, was this for speaking? a. Yes b. No

22. If you are familiar with a certain rating scale from a commercially available test or that is in use at your institution, please specify the name of the test or scale.

Test(s): _____; _____
(e.g., *Test of Spoken English*)

Appendix C: Instructions and Practice Items for Verbal Protocol Condition, Study 2

The purpose of this study is to investigate how listeners perceive non-native speech, particularly with respect to pronunciation. Your role is of listener, rater, and articulator of your thoughts. Your task is as follows:

1. Listen

You will listen to a 20 second speech sample of a non-native English speaker telling a picture story. The first language of the speaker will be either Russian or Mandarin Chinese.

2. Rate

You will rate the speaker for pronunciation on the three separate rating scales provided on page 3 (for comprehensibility, accentedness, and fluency).

3. Think aloud

When the researcher pauses the tape, you will be asked to think aloud, that is, to articulate your thought processes as you decide what scores to assign the speaker.

NOTE: It is up to you whether you rate the speaker while the speech sample is still playing or once the speech sample is over. In either case, your job is to verbalize what you are thinking or remember thinking when making your decision about what scale level to assign on the rating scales provided.

The most important part of the task is for you to say everything that you are thinking or remember thinking out loud. If you are silent for any period of time, the researcher will prompt you to keep talking.

Tips:

- Try not to analyze your thoughts or provide explanations for them – it is important for you to simply report what you are/were thinking while listening and scoring.

- Don't try to censor aspects of the listening/rating process that you perceive to be irrelevant or judgmental – give an honest account of your thoughts.

Below are the three 5-point rating scales that you will use to evaluate each speaker for the following aspects of speech related to pronunciation: comprehensibility, accentedness, and fluency. Use the definitions below to guide your judgments.

1. **Comprehensibility** – How easy you think the speaker is to understand.

2. **Accentedness** – How different you think the speaker sounds from a native speaker of North American English

3. **Fluency** – How smooth you think the speaker's oral delivery is based on his/her use of pauses, hesitation markers, fillers (e.g., um, uh), etc.

Now, let's try a few examples so you can familiarize yourself with the task. Please circle one number only for each of the three scales in addition to articulating your thoughts about the process. It is important for you to ask if you have any questions.

Note: Try to use the whole rating scale.

Example 1

1. Comprehensibility	1	2	3	4	5
	Very hard to understand			Very easy to understand	
2. Accentedness	1	2	3	4	5
	Heavily accented			Not accented at all	
3. Fluency	1	2	3	4	5
	Very dysfluent				Very fluent

Example 2

1. Comprehensibility	1	2	3	4	5
	Very hard to understand			Very easy to understand	
2. Accentedness	1	2	3	4	5
	Heavily accented			Not accented at all	
3. Fluency	1	2	3	4	5
	Very dysfluent				Very fluent

Appendix D: Post-Task Interview Guidelines, Study 2

1. How did you find thinking aloud? Was it easy/difficult, natural/unnatural?

    - Do you think articulating your thoughts affected your ratings? If so, how?
    -Think aloud twice removed vs. once removed?

2. Do you recall any specific difficulties while you were rating?

    - Were there particular points in the scale where you found it more difficult to assign scores than others?

3. At what moment did you typically make your decision about what score to assign?

    - By the time the recording was over, were you already finished making your decision about what score to assign?
    - Did you find certain speakers to be easier to score than others? Why?

4. How did you find the rating scales?

    - If there had been rating scale descriptors, would this have facilitated or complicated the task for you?
    - Rating experience? Influence of previous rating scales?

5. What criterion would you say was the most important for you overall? What distinguished a high-rated speaker from a low-rated speaker in general?

    - For comprehensibility? For accentedness? For fluency?
    - Did your ratings usually match up? For example, if you gave a speaker a 6 for comprehensibility, did you tend to assign the same score for accentedness and fluency?

6. Do you have any other comments about the speech samples that you rated today?

    - Is there anything else you'd like to talk about with regards to the ratings, speech samples or what I've asked you to do today?

Appendix E: Instructions for Teacher Written Reports and Practice Item, Study 3

The purpose of this study is to investigate the factors that ESL teachers find most salient when listening to and scoring L2 accentedness and comprehensibility. Your task is as follows:

Listen

You will listen to a short speech sample of an adult Francophone speaker telling a picture story in English. You may listen to the speech as many times as you require. You can play, stop, or rewind the recording using the computer mouse. This is a self-paced task.

Rate

Once you have finished listening to the speech, you will pause the recording and will use the provided Microsoft Word document to rate the L2 speaker for accentedness or comprehensibility using a 9-point scale.

It is your choice whether you rate accentedness or comprehensibility first. This will probably depend on which scoring decision you arrive at most quickly.

Type in your impressions of the most salient aspects of the speech that you took into account when rating

Directly under the scale where you provided your rating, you will type in the aspects of the speech that most affected your scoring decision directly into the text box. These can be in point form.

What we're interested in here are the most striking aspects of the speech from your perspective that influenced your rating. Anything related to the speaking style that factored into your scoring decision for accentedness or comprehensibility is relevant.

If you happened to notice or were distracted by some aspect of the speech but decided NOT to take that into account in your scoring, please convey this in your comments. Just say that you noticed it but decided it shouldn't influence your ratings.

Similarly, if you experienced a dilemma when assigning a score, please describe this and how you arrived at your rating decision. If you didn't find anything noteworthy about the speech, you can just state that too.

At times, you may feel that your thoughts about the speech are not relevant to our study. Rest assured that they probably are. Try not to censor these thoughts. Instead, let your

fingers do the typing and simply report what comes into your head as your reflect on the speech and the score you assigned for each measure.

<u>Provide your second rating and comments</u> by following the same procedure

Once you have completed these steps, you will conduct your second rating for whichever measure you still have to rate (either accentedness or comprehensibility) by following the same procedure.

Rating scales

You will use the 9-point rating scales shown below to evaluate each speaker for accentedness and comprehensibility. Please use the definitions below to guide your judgments.

Comprehensibility – How easy you think the speaker is to understand.

Accentedness – How different you think the speaker sounds from a native speaker of North American English.

To indicate your score, please put an X in the box with the rating that best corresponds to your opinion. In the example below, a speaker is rated for comprehensibility at 7.

<u>EXAMPLE</u>:

COMPREHENSIBILITY:

1	2	3	4	5	6	7X	8	9

1 = hard to understand 9 = easy to understand

Comments about comprehensibility

In this textbox, you will type in the aspects of the speech that you found most striking and took into account when rating comprehensibility.

Appendix F: Teacher Post-Rating Questionnaire, Study 3

NOTE: Only questions 3 and 4 are reported on in Study 3

Summarizing your listening/rating experience

1. Which of the following aspects of the speech do you think affected the ratings you assigned for accentedness and comprehensibility? Please indicate which of the below were relevant to your scoring decisions by typing an 'X' next to that aspect for accentedness and comprehensibility.

If the aspect of speech that is listed did not affect your ratings, please leave it blank. This list is not comprehensive, so if there are other relevant aspects that do not appear here, please add them to the bottom of the list and then indicate whether they are important for accentedness or for comprehensibility by placing an 'X' in the appropriate column.

Aspects of speech	Accentedness	Comprehensibility
1. Lexical errors		
2. Richness of vocabulary		
3. Grammatical errors		
4. Pronunciation of vowels or consonants		
5. Word stress (emphasis of most prominent syllable in a word)		
6. Speaker's storytelling ability		
7. Story cohesion (flow of ideas, etc.)		
8. Intonation		
9. Unnecessary repetition of words or syllables		
10. Number of "ums" and "uhs"		
11. Number of silent pauses		
12. Production of fluent stretches of speech		
13. Natural sounding rhythm		
14.		
15.		
16.		

2. Of the aspects you put an 'X' next to, please rank order the top 3 that you felt were most important to your decision making in the table below. You can do so by typing in the number that is listed next to the aspect of speech. For example, if 'lexical errors' was the most important aspect for accentedness, you would type '1' into the first row under

'accentedness.' If the first aspect you added to the list was the most important, you would type in '14.'

Influences on your ratings	For accentedness	For comprehensibility
Most important aspect of speech		
Second most important		
Third most important		

3. For this study, we defined comprehensibility as "how easy you think the speech is to understand." This definition is pretty broad. Which of the below was closest to your interpretation of comprehensibility? Please **bold** 'a,' 'b,' or 'c' to indicate you choice.

(a) how easy the individual *words* that the speaker articulated were to understand
(b) how easy the speaker's *story* was to understand
(c) other (please specify):

4. If you have any other comments about your listening and rating experience today, please share them with us: