



Title: Exploring the influence of suprasegmental features of speech on rater judgements of intelligibility

Name: Thomas Michael Rogers

This is a digitised version of a dissertation submitted to the University of Bedfordshire.

It is available to view only.

This item is subject to copyright.

EXPLORING THE INFLUENCE OF SUPRASEGMENTAL FEATURES OF SPEECH ON RATER JUDGEMENTS OF INTELLIGIBILITY

A thesis submitted to the University of Bedfordshire in partial fulfilment of
the requirements for the degree of Doctor of Philosophy

by

Thomas Michael Rogers

Centre for Research in English Language Learning and Assessment

University of Bedfordshire

January 2018

Declaration

I, Thomas Rogers, declare that this thesis and the work presented in it are my own and have been generated by me as the result of my own original research.

*Exploring the Influence of Suprasegmental Features of Speech on Rater Judgements of
Intelligibility*

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have cited the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission.

Signed:

Date:

Abstract

The importance of suprasegmental features of speech to pronunciation proficiency is well known, yet limited research has been undertaken to identify how raters attend to suprasegmental features in the English-language speaking test encounter. Currently, such features appear to be underrepresented in language learning frameworks and are not always satisfactorily incorporated into the analytical rating scales that are used by major language testing organisations. This thesis explores the influence of lexical stress, rhythm and intonation on rater decision making in order to provide insight into their proper place in rating scales and frameworks.

Data were collected from 30 raters, half of whom were experienced professional raters and half of whom lacked rater training and a background in language learning or teaching. The raters were initially asked to score 12 test taker performances using a 9-point intelligibility scale. The performances were taken from the long turn of Cambridge English Main Suite exams and were selected on the basis of the inclusion of a range of notable suprasegmental features. Following scoring, the raters took part in a stimulated recall procedure to report the features that influenced their decisions. The resulting scores were quantitatively analysed using many-facet Rasch measurement analysis. Transcriptions of the verbal reports were analysed using qualitative methods. Finally, an integrated analysis of the quantitative and qualitative data was undertaken to develop a series of suprasegmental rating scale descriptors.

The results showed that experienced raters do appear to attend to specific suprasegmental features in a reliable way, and that their decisions have a great deal in common with the way non-experienced raters regard such features. This indicates that stress, rhythm, and intonation may be somewhat underrepresented on current speaking proficiency scales and frameworks. The study concludes with the presentation of a series of suprasegmental rating scale descriptors.

Acknowledgements

Several people were involved in providing access to the test taker responses necessary to carry out this study. I would like to thank Jaime Dunlea, Fiona Barker, Ardeshir Geranpayeh, Bruce Howell, John Slaght, Sarah Brewer, and Veronica Benigo for their generosity during this project. I am also grateful to the panel of phoneticians who helped me make sense of the data: Yuni Kim, Richard Cauldwell, Linda Shockey, and Rachel Smith. Financial support was provided by Pearson, for which I am grateful.

Of course this research would have been impossible without the participants who kindly donated their time. I am also indebted to Stefan O'Grady for discussion and comments on drafts of chapters, and Dan Frost and Jane Setter for helpful discussion in the early stages of the project.

The staff at CRELLA have been a regular source of support throughout this project, primarily John Field who supervised my doctoral studies. I greatly appreciate the guidance and mentorship that he provided. My sincerest thanks also go to Fumiyo Nakatsuhara who provided insightful feedback at critical junctures.

Finally, my most heartfelt gratitude goes to my parents, and to Cara and Masha for their loving support.

This thesis reports on research using examination data provided by Cambridge English Language Assessment.

Table of Contents

Abstract.....	v
Acknowledgements	vi
Table of Contents	vii
List of Tables	xi
List of Figures.....	xii
Additional Material	xiii
Abbreviations and Acronyms	xiv
The International Phonetic Alphabet.....	xv
Chapter 1: Introduction.....	1
1.1 Rationale for the Study	3
1.2 Setting for the Research	8
1.2.1 Current Practice in Pronunciation Assessment.....	8
1.2.2 The Context for this Research.....	12
1.3 Significance of the Study	13
1.4 Important Terms Defined	14
1.5 Outline of the Thesis.....	15
Chapter 2: Literature Review.....	21
2.1 Suprasegmental Features of Speech.....	22
2.1.1 Stress.....	23
2.1.2 Intonation	26
2.1.3 Rhythm	29
2.2 Defining Intelligibility	32
2.3 Suprasegmental Correlates of Intelligibility	34
2.3.1 Lexical Stress.....	35
2.3.2 Intonation	36
2.3.3 Rhythm	42
2.3.4 Rater Awareness of Suprasegmental Features	44
2.4 Frameworks and Rating Scales	49
2.4.1 Current Frameworks	50
2.4.2 Rating Scale Development and Validation.....	55
2.4.3 Current Rating Scales	60
2.5 Raters.....	63

2.5.1 Influence of Rater Experience	64
2.5.2 Linguistic Background	67
2.6 Summary, Research Questions and Hypotheses.....	70
2.6.1 Research Questions.....	71
2.6.2 Original Contribution to the Literature.....	77
Chapter 3: Research Design	79
3.1 Method	80
3.1.1 Methodological Approach.....	81
3.1.2 Ratings Scale.....	83
3.1.3 Retrospective Verbal Reports.....	85
3.1.4 Analysis	87
3.2 Participants.....	92
3.3 Materials.....	97
3.3.1 Audio Samples.....	97
3.3.2 Rating Sheet	103
3.3.3 Procedure	105
3.4 The Pilot	109
3.5 Ethical Considerations	110
3.6 Summary	111
Chapter 4: Quantitative Findings and Discussion	113
4.1 Quantitative Analysis	113
4.2 Quantitative Findings.....	114
4.3 Quantitative Discussion	132
4.3.1 Consistency within the Experienced Rater Group [RQ1]	133
4.3.2 Influence of Rating Experience [RQ2]	136
Chapter 5: Qualitative Findings and Discussion	145
5.1 Qualitative Analysis	145
5.1.1 Transcription	145
5.1.2 Coding	146
5.1.3 Categorising Codes.....	148
5.1.4 Analysis of Categories	151
5.2 Qualitative Findings	151
5.2.1 Experienced Raters.....	151

5.2.2 The Panel of Phoneticians	164
5.2.3 Non-Experienced Raters	165
5.2.4 Rater Familiarity	171
5.2.5 Summary of Major Qualitative Findings	173
5.3 Qualitative Discussion	174
5.3.1 Summary of Qualitative Discussion	188
Chapter 6: Towards Suprasegmental Rating Scale Descriptors	191
6.1 Approach to Developing Scale Descriptors	191
6.2 Suprasegmental Rating Scale Descriptors	197
6.2.1 Descriptor Development	200
6.2.2 Descriptor Quality	203
6.3 Discussion	206
6.3.1 Reference to Existing Scales and Frameworks	209
Chapter 7: Conclusion	217
7.1 Summary of Findings	218
7.1.1 RQ1	219
7.1.2 RQ2	219
7.1.3 RQ3	220
7.2 Applications	221
7.3 Limitations of the Study and Future Directions	225
7.3.1 Trialling Suprasegmental Descriptors	228
7.4 Concluding Remarks	229
References	231
Appendices	259
Appendix A: Pronunciation Frameworks	259
Appendix B: Speaking Rating Scale Descriptors	263
Appendix C: Rating Sheet	269
Appendix D: Sample Interview Transcriptions	270
Appendix E: Preliminary Audio Rationales	285
Appendix F: Final Audio Rationales	302
Appendix G: Standardised Scores	303
Appendix H: Facets Output	304
Appendix I: Raw Scores	309

Appendix J: Categories and Codes310
Appendix K: Integrated Display312

List of Tables

Table 2.1: Lists of Suprasegmental Features	22
Table 3.1: Summary of Raters	95
Table 3.2: Rating Experience	97
Table 3.3: Summary of Audio Stimuli	102
Table 3.4: Rating Scale Criteria	104
Table 4.1: Test Taker Summary Statistics	118
Table 4.2: Average Intelligibility Estimates	119
Table 4.3: Rasch-Andrich Thresholds	120
Table 4.4: Category Outfit Mean-Square	123
Table 4.5: Rater Performance by Group	124
Table 4.6: Rater Fit Statistics (All Raters)	125
Table 4.7: Rater Performance (All Raters)	127
Table 4.8: Rater Summary Statistics	127
Table 4.9: Test Taker Fair Average Scores by Exam	128
Table 4.10: Single Rater-Rest of Raters Correlation within Groups	130
Table 4.11: Rater Group Severity	131
Table 5.1: Coding Categories	149
Table 5.2: Instances of Rhythm and Strain Coded Together	153
Table 5.3: Number of Times Categories Mentioned by Experienced Raters	158
Table 5.4: Rhythm Codes used by Experienced Raters	159
Table 5.5: Intonation Codes Assigned by Experienced Raters	161
Table 5.6: Number of Times Categories Mentioned by Rater Group	165
Table 5.7: Number of Times Categories Mentioned by Non-Experienced Raters	167
Table 6.1: Test Takers Grouped by Fair Average Score	193
Table 6.2: Scale Groups	196
Table 6.3: Suprasegmental Rating Scale Descriptors	198
Table 6.4: Lexical Stress Descriptors	205

List of Figures

Figure 2.1: The 2001 CEFR Phonological Control Grid	52
Figure 2.2: Horner's Proposed Phonological Control Grid	53
Figure 3.1: Overview of the Design	79
Figure 3.2: Logit Difference Between Test Taker Ability and Item Difficulty Against Probability of Success	88
Figure 4.1: Intelligibility Variable Map	115
Figure 4.2: Experienced Rater Scores by Audio Clip	116
Figure 4.3: Non-Experienced Rater Scores by Audio Clip	117
Figure 4.4: Scale Category Probability for All Raters	121
Figure 4.5: Fair Average Scores by Exam	129
Figure 5.1: Percentage of Experienced Raters Mentioning Rhythm	152
Figure 5.2: Percentage of Experienced Raters Mentioning Intonation (ordered by fair average scores)	155
Figure 5.3: Percentage of Experienced Raters Mentioning Lexical Stress	156
Figure 5.4: Intonation Codes by Rater Group	169
Figure 6.1: Experienced Rater Intelligibility Variable Map	194
Figure 6.2: Cambridge Examinations and the CEFR	195

Additional Material

CD Track List (see the inside back cover of this thesis)

1. Speaker 1 (00:59)
2. Speaker 2 (00:56)
3. Speaker 3 (00:50)
4. Speaker 4 (01:01)
5. Speaker 5 (00:57)
6. Speaker 6 (00:48)
7. Speaker 8 (00:56)
8. Speaker 9 (00:56)
9. Speaker 10 (00:58)
10. Speaker 11 (00:59)
11. Speaker 12 (00:57)
12. Speaker 14 (01:10)

Abbreviations and Acronyms

ACTFL	American Council on the Teaching of Foreign Languages
ALTE	Association of Language Testers in Europe
CAE	Cambridge English: Advanced
CEFR	Common European Framework of Reference for languages
CPE	Cambridge English: Proficiency
EFL	English as a Foreign Language
ESL	English as a Second Language
ETS	Educational Testing Service
ESL	English as a Second Language
FCE	Cambridge English: First
GRM	Grounded Theory Method
GSE	Global Scale of English
IELTS	International English Language Testing System
IPA	International Phonetic Alphabet
IRT	Item Response Theory
KET	Cambridge English: Key
L1	First Language
L2	Second Language
MFRM	Many-Facet Rasch Measurement
PET	Cambridge English: Preliminary
PTE-A	Pearson Test of English: Academic
TESOL	Teaching English to Speakers of Other Languages
TOEFL iBT	Test of English as a Foreign Language internet based test

The International Phonetic Alphabet

CONSONANTS (PULMONIC)

© 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			ʀ					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

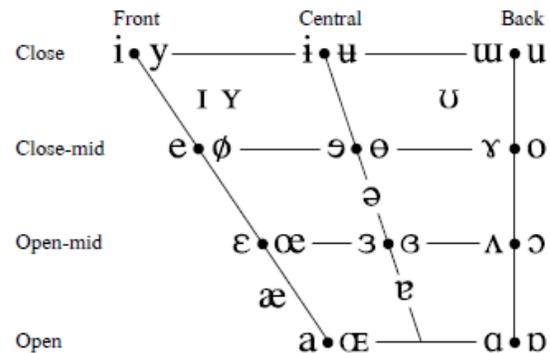
CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
◌ ɸ Bilabial	ɓ Bilabial	ʼ Examples:
◌ ɸ Dental	ɗ Dental/alveolar	ɸ' Bilabial
◌ ɸ (Post)alveolar	ɟ Palatal	t' Dental/alveolar
◌ ɸ Palatoalveolar	ɠ Velar	k' Velar
◌ ɸ Alveolar lateral	ɠ Uvular	s' Alveolar fricative

OTHER SYMBOLS

ɱ Voiceless labial-velar fricative	ç ʝ Alveolo-palatal fricatives
ʋ Voiced labial-velar approximant	ɻ Voiced alveolar lateral flap
ɰ Voiced labial-palatal approximant	ɥ Simultaneous ʃ and x
ħ Voiceless epiglottal fricative	
ʕ Voiced epiglottal fricative	Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.
ʔ Epiglottal plosive	

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

ts̺ k̟p̟

International Phonetic Association (2017)

Chapter 1: Introduction

The purpose of a language test is to measure a test taker's knowledge or proficiency in a given language. Such tests are employed in educational establishments as well as workplaces and dedicated testing centres, and can be employed for relatively low-stakes purposes, such as classroom progress assessment, or for high-stakes purposes, such as for university entry, professional progression, or emigration. Language tests often include separate assessments of reading, writing, listening and speaking, or a combination of these. Speaking is an important component in this because many of the reasons for taking a test relate to being able to communicate orally in the given language. For example, in a professional context test takers might need to hold meetings in the target language, and in a university they may be expected to deliver presentation. Within the assessment of speaking there are a series of proficiencies. In addition to pronunciations, rating scale criteria typically assess: 1) a test taker's control over grammar and vocabulary, sometimes referred to as language use; 2) criteria such as fluency, coherence, and discourse management, which can be described as delivery; and 3) aspects of how well the test taker achieves the goals of the communicative interaction. Pronunciation proficiency is a primary focus of speaking assessment because it goes to the heart of what oral communication requires. That is, the speaker's ability to produce words and utterances which a listener can regard as intelligible.

The general purpose of this research was to better understand the role of suprasegmental features of speech in rater judgements of intelligibility in English language speaking tests. The motivations for this were threefold: 1) there is a relative lack of detail in reference to these features on certain scales and language frameworks that are currently in use; 2) research examining the link between suprasegmental features of speech and pronunciation judgements is somewhat patchy, with contrasting findings in some areas and complete gaps in others; and 3) over a number of years as an examiner for English speaking tests, the researcher developed a curiosity relating to the capacity of raters to consistently detect and interpret suprasegmental features of speech in language assessments. Several studies have addressed this last issue tangentially, but to date there has not been any research which focuses primarily on the suitability of professional raters to consistently interpret and judge the suprasegmental element of test taker performances. The ultimate aim of this project was to provide enough evidence to produce a series of suprasegmental rating scale descriptors which could be employed to supplement the descriptors currently used.

Tests of spoken English are delivered in a range of formats. They can incorporate monologues or dialogues or a combination of the two. For example, candidates taking the Cambridge English: First (FCE) exam have to respond to an interlocutor's questions, as well as have a discussion with their partner, and deliver a monologue by describing a picture. In most tests, candidates are required to answer questions or respond to written, verbal or graphical prompts. For instance, in the Pearson Test of English: Academic (PTE-A) test takers hear a lecture and then summarise it, and in Cambridge English: Advanced (CAE), they look at some pictures and describe them.

Scoring presents a challenge for speaking tests. The assessment can take place in a face-to-face format in the presence of one or more examiners, or can be recorded and either sent to an examiner or scored by computer. The most common format is for a test taker to perform a

communicative task in the presence of a rater who compares their performance to a rating scale in order to establish the test taker's level of proficiency. Rating scales describe the construct that raters are required to assess, and rater training is employed to ensure that raters can match performances to that construct in a consistent way. This is a necessarily subjective form of assessment because speaking proficiently requires the complex interplay of a wide range of competencies. Rigorous training of raters and the careful development of rating scales can bring a degree of objectivity to these assessments, but rating scales take a range of forms with varying advantages and disadvantages. In high-stakes testing, analytical scales are common. They describe the performance at each proficiency level in detail, often over more than one criterion. Scales can also be holistic, however, giving raters a broad direction as to the overall performance at each level.

This introductory chapter outlines the rationale for the project, and summarises the way pronunciation assessment is currently undertaken. The context of the research is then described, after which the significance of this study is examined. The final two sections consist of a definition of the important terms, followed by a summary of the contents of this thesis.

1.1 Rationale for the Study

Understanding what happens in the speaking test encounter is important because such tests can be high stakes. High-stakes pronunciation testing has been traced as far back as a biblical reference to a test where failure resulted in death (Isaacs & Harding, 2017). Today, high-stakes testing is often a necessary and important part of establishing whether applicants will have the required language skills to be successful in a range of contexts, such as on English-medium university courses, and to establish whether they can operate safely in certain work environments.

Assessing spoken English has unique challenges. The ephemeral nature of speech and the variation of norms and standards, even within speakers of the same language community, makes defining proficiency levels difficult. A complicating factor is that raters seem to prefer scales that are as precise as possible (Isaacs, Trofimovich, Yu, & Chereau, 2015) and fine-grained detail of pronunciation performance is sometimes lacking on speaking scales. This may result in raters being unable to assign scores with the maximum degree of precision, but it also represents a risk to construct validity in the sense that the rating scale represents the speaking construct as it is practically applied by raters. This creates a tension between, on one hand, raters' desire for detailed scales and the need for test developers to define the construct they expect raters to measure, and on the other hand, the variation in norms that exist in English pronunciation. Increasing the level of detail in rating scales therefore requires a validity argument that grounds the scale in empirical research or expert judgement. For the sake of reproducibility and fairness, such a scale must be consistently applicable by raters, and this is the primary difficulty. In the USA, where much early language testing research took place, oral proficiency assessments were avoided by language testing practitioners until 1926 because rater agreement, a critical element of test validation, was difficult to obtain (Fulcher, 2003). Pronunciation descriptors must be detailed enough to adequately reflect differences in proficiency, but broad enough for raters to be capable of applying them consistently. The research presented here investigates this fine balance between precision and reliability.

Pronunciation is one criterion against which speaking proficiency is usually assessed and scales vary in the degree of detail they provide for pronunciation descriptors. There is a clear spectrum of scales, from those that make no mention of stress, rhythm, and intonation, such as the public version of the IELTS speaking descriptors (IELTS, 2012: 19), to those that address such features to a moderate degree of detail, such as the PTE-A oral fluency and pronunciation scales. The issue is particularly marked in language frameworks, such as the CEFR, which also suffer from a

underspecified approach to pronunciation. The 2001 version of the CEFR (Council of Europe, 2001) was criticised for its inconsistent and problematic approach to pronunciation, although this was rectified to a great degree by the 2017 revisions (Council of Europe, 2017). Despite this improvement, one thing language frameworks have in common is a lack of precision in reference to these features.

It is possible to get a sense of why suprasegmental features may be somewhat less well represented in certain scales and frameworks by examining existing research in the field. Phoneticians have identified and described suprasegmental features objectively and phonologists have demonstrated the communicative value of these features to native and non-native English speakers. Yet as it stands, there is no clear hierarchy for the acquisition of suprasegmental features (Horner, 2014: 111). This presents a challenge to developing logical, usable scales and frameworks. Assessment researchers have been able to correlate such features to scores on certain oral proficiency criteria, for instance Kang et al. (2010) identified the impact of pitch choice on oral proficiency to a fine degree of detail, but such work is limited by the assessment context which requires scale descriptors to be reliably interpretable and applicable by raters. Therefore, the issue goes beyond identifying the link between the features present in the speech signal and the resulting scores because of the necessary process of interpretation required of raters. There is a complex interaction between what is objectively present in a test taker's speech signal, the way it is perceived by a rater, and the role that a rater's interpretation plays in selecting the resulting score.

Raters appear to consider a wide range of features when making their judgements. For instance, Hayes-Harb and Hacking (2015: 43) discovered that "a straightforward bottom-up analysis of the speech signal" was insufficient to explain the varying ways raters made judgements of accent. The role of suprasegmental features in this interaction is not well understood. This is particularly

important because features such as intonation and rhythm are regarded as fundamental to the organisation of speech and critical to resolving ambiguities (Beckman, 1996). The purpose of this research is to investigate the extent to which suprasegmental features of speech provoke inconsistency in rater interpretations.

Assessment researchers often use quantitative approaches to establish the quality of scoring, and such an approach is sensible given that the outcome of speaking assessment is usually a score. However, rater effects are not always clear in scores alone. The influence raters have on scores might only be observable by speaking to them (see, for instance, Orr, 2002). This explains why understanding what raters attend to when making judgements is essential for development of assessment criteria and rating scales (Taylor & Galaczi, 2011: 204).

Although some studies have demonstrated that raters largely agree on the features of speech influential to scores (e.g. in a paired oral task, Ducasse & Brown, 2009: 440), others tend to indicate that different raters approach rating from different perspectives (Orr, 2002). Even when raters are provided with a detailed rating scale, wide variations in performances can result in similar scores (e.g. Douglas & Selinker, 1992; Douglas & Chapelle, 1993). Suprasegmental features have been examined in studies addressing what raters attend to when making judgements but it is still not clear the extent to which raters approach suprasegmental features consistently, and precisely how their attention to specific features relates to their scoring patterns. This is critical for the consistent application of scales and therefore the validity of an assessment. A good illustration of this difficulty is a study by May (2011), who found that raters mentioned features of speech that were not included in the rating scale criteria when reporting their rating processes. This suggests that the ideal target construct represented by the rating scale was not the actual construct being operationalised by raters. In other words, the raters did not judge test takers solely against the criteria they are presented with, rather they engaged their own unique

perspective on the performance. Given that suprasegmental features are somewhat neglected on current rating scales, this leaves much scope for raters to engage their own perception of the influence of such features on scores. Indeed, Levis (2006) argues that whether pronunciation appears on scales or not, it will be used by raters when making their judgements, possibly becoming “a source of unsystematic variation” in the test as a result (p245).

Potentially a major factor influencing what features of speech raters can consistently attend to is rater experience. Saito et al. (2016: 143) stated that the influence of experience profile on the speech characteristics that raters heed is unknown. Although the impact of accent familiarity on scores is receiving more attention (e.g. Browne, 2016), the role of interpretation of specific features of speech and their impact on judgements in relation to rating experience is under-examined. This presents a risk of variation among raters as they often come from a variety of linguistic backgrounds, but it also represents a potential difference between the listening that takes place in the assessment context and the listening that often takes place in the real world. The implications of this are that the ability of a rater to decode a test taker’s speech signal may be wholly different to the way a typical listener does the same thing. Someone taking an exam for migration to the UK, for instance, is unlikely to encounter listeners in their day-to-day life who are as linguistically sophisticated as those who judge their performance in the exam. It is important, therefore, to establish the degree to which the listening taking place in the assessment is consistent with that of real-world contexts.

In summary, it is not clear what raters notice when they make their pronunciation judgements. There is some evidence that they do attend to suprasegmental features of speech, but it is not clear what importance they give to such features. If there is evidence that raters do attend to suprasegmental features consistently then it may be possible to make relevant additions to pronunciation rating scale descriptors.

1.2 Setting for the Research

The setting for this research is the English language test and, specifically, speaking assessments. The mode of assessment and the types of tasks that can be included in a speaking test vary greatly, so this section reports two case studies which illustrate the current state of speaking assessment. Following that, the specific context for the materials used for data collection in this research is described.

1.2.1 Current Practice in Pronunciation Assessment

The two speaking tests described here are the speaking paper of the *Cambridge English: First* exam, and Part 1 of the *Pearson Test of English: Academic*. These exams are used for making a wide range of high-stakes decisions, including university entry, professional entry or progression, and migration. As such, it is critical that these exams produce valid, fair outcomes.

These exams were selected for review here because they have different modes of delivery, item types, and scoring, and therefore reflect the range of high-stakes speaking tests that are available to test takers. The role of suprasegmental features of speech in the scales used in these exams is examined more closely in Chapter 2.

Cambridge English: First (FCE)

This exam is designed for test takers who are aged 13 – 18 (Cambridge Assessment, 2017c) and the purpose is to assess whether candidates are able to live and work independently in an English-speaking environment. The exam targets level B2 on the Common European Framework of Reference for Languages (CEFR), which means test takers should be able to “interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party” (Council of Europe, 2001: 24).

FCE consists of four papers, one of which is a speaking paper. This paper is delivered face-to-face in the presence of two examiners: an assessor, who scores the performances on an analytical scale; and an interlocutor who interacts with the test takers and times the exam, as well as provides a score based on a global scale.

Candidates take the speaking test in pairs or groups of three. The exam takes 14 minutes and consists of four parts. In the first part, the interlocutor introduces the examiners and asks each test taker some standard introductory questions. These include things like “how do you spend your evenings?” and “what did you do on your last birthday?” (Cambridge Assessment, 2017b). This is an interaction between the interlocutor and each test taker individually, and it requires candidates to express their opinions. In part two of the exam, test takers are presented with some images and are asked to compare the photographs and discuss a particular topic in relation to them. For instance, the first test taker could receive a photograph of an injured sports person and of someone asking a police officer for directions and be provided with the question “how important is it to help people in these situations?” (Cambridge Assessment, 2017b). After talking about the photographs for one minute, the other test taker is asked questions about them. This section of the exam is designed to give each test taker the opportunity to deliver a monologue.

Part three is a discussion between the test takers. They are given a topic and a diagram which contains some ideas about how they might address that topic. They are given 15 seconds to independently think about how to respond to the task and then three minutes to talk to each other. This part of the exam assesses the interaction of test takers and their ability to negotiate towards a conclusion. Finally, part four takes the remaining four minutes of the test time and consists of a discussion between the test takers in response to questions asked by the interlocutor, such as “why do you think people like to go away on holiday?” (Cambridge Assessment, 2017b).

Each test taker's performance is scored on a 6-point holistic global achievement scale by the interlocutor, and against four criteria on a 6-point analytical scale by the assessor. Although full descriptors are not publicly available, an indication of the method of assessment can be obtained from the teacher's handbook (Cambridge English, 2016b). The global scale includes descriptors such as "handles communication in everyday situations, despite hesitation" (p82). The criteria that the assessors judge test takers against are grammar and vocabulary, discourse management, pronunciation, and interactive communication. The pronunciation descriptors include terms such as "sentence and word stress is accurately placed" (p82).

The scoring process results in test takers receiving five scores for their performance, one of which directly targets pronunciation, and one of which incorporates pronunciation in global terms. These scores are combined and make up 20% of the total score that a test taker receives for the examination, the remaining 80% being made up of the listening, reading, and writing papers.

Pearson Test of English: Academic (PTE-A)

PTE-A is an exam primarily aimed at test takers who are trying to gain entry to universities that have an English language requirement, although it is also taken for professional progression and migration (Pearson, 2017e). The exam is administered by computer. Test takers read and listen to prompts which are presented on a computer, and they respond to these prompts by speaking into a microphone. The responses are recorded and scored using an automated scoring system.

The test consists of three parts: speaking and writing, reading, and listening. The speaking and writing part consists of eight sections for which test takers have up to 93 minutes. The first section is not scored, and simply allows test takers to record messages for institutions that will receive

their scores. The second and third sections test the ability to read aloud and repeat an utterance. Test takers respond to prompts by speaking into a microphone. In the second section, test takers are presented with a short text of up to 60 words. They have 30 – 40 seconds to prepare and then must read the text aloud. This assesses test taker ability to identify a writer’s tone or attitude and communicate this effectively orally. In the third section, they hear a short phrase, such as “it is a good idea to start your dissertation with a review of the literature” (Pearson, 2017b), and then must repeat it. Among other things, this section assesses a test taker’s ability to articulate words and utterances appropriately.

In the fourth section, test takers are presented with an image, for example, a map of Africa which shows gorilla distribution in each country (Pearson, 2017c). They are then allowed to prepare a response before describing the image in as much detail as possible. In the fifth section, test takers listen to 90 seconds of a lecture on a topic, such as art restoration. They also see an image relevant to the topic. They can take notes during the task and after the lecture they have 10 seconds to prepare before re-telling what they heard. As well as testing listening and notetaking, this part of the test assesses the ability of test takers to speak for a specific purpose, as well as organising their discourse logically. In section six test takers hear a short question, such as “what type of periodical is published on a daily basis?” (Pearson, 2017d). After hearing the question, test takers have 10 seconds to answer it. The final two sections are devoted to writing rather than speaking.

Each task is scored on two 6-point scales: a pronunciation scale which includes descriptors such as “stress-dependent vowel reduction may occur on a few words” (Pearson, 2017a: 24), and an oral fluency scale which includes descriptors such as “speech has an acceptable rhythm with appropriate phrasing and word emphasis” (Pearson, 2017a: 24). Notably though, the scores are applied by computer rather than by a human. The scores for each section are combined and

reported on a score report. This report consists of an overall score, a score for each skill (listening, reading, speaking, and writing), and six scores for enabling skills, which includes one for pronunciation (Pearson, 2017a).

1.2.2 The Context for this Research

The context for this research is international English language speaking assessments. As this research is primarily interested with the way human judges make their decisions, and since the scores assigned by computer scored speaking tests are derived from human scored judgements, it was decided to collect data using material from tests which are directly human scored, rather than tests where the scores are applied by computer.

The material used for collecting data was provided by Cambridge Assessment. This material constituted 12 audio recordings of face-to-face speaking exams that had taken place in the UK. The exams included were Cambridge English: Preliminary (PET), Cambridge English: First (FCE), and Cambridge English: Advanced (CAE), which cover approximately levels B1 – C1 on the CEFR. PET and FCE are designed for learners who are aged approximately 11 – 18, whereas CAE is designed for older test takers (Cambridge Assessment, 2017c).

Cambridge Assessment employs thousands of examiners, all of whom are required to have substantial English language teaching experience (Cambridge Assessment, 2017a). A sample of examiners (N = 15) were recruited to judge the recordings of authentic speaking exams and provide intelligibility scores, based on a 9-point intelligibility scale, and verbal reports. The scores reflected the raters' perceptions of how intelligible each speaker was, and the verbal reports reflected the features of speech they attended to when making their judgements. A sample of

non-experienced examiners (N = 15) were also recruited to examine the extent to which the experienced raters' perceptions were consistent with those of typical listeners.

1.3 Significance of the Study

The high stakes nature of much pronunciation testing, and the way in which rating scale criteria temper the subjective nature of oral assessment, means that researching the efficacy of rating scales is an important activity. Oral proficiency interviews have been a subject of research since at least the Second World War (Chalhoub-Deville & Fulcher, 2003) and the use of rating scales can be traced as far back as 1864 (Chadwick, 1864, cited in Fulcher, 2015), but this study is timely because until recently there had been limited research into the role of pronunciation and even less so on the role of suprasegmental descriptors. This thesis comes at a time of greater calls for mixed-methods studies examining the role of suprasegmental features on rating scales (e.g. Galaczi, Post, Li, Barker, & Schmidt, 2016: 179), and at a time when such descriptors are beginning to be examined more thoroughly (e.g. Frost & O'Donnell, 2015; in press). Concluding their volume on second language pronunciation, Isaacs and Trofimovich (2016) present several questions that they consider to be important for L2 pronunciation assessment in future, one of which is: which pronunciation features should be prioritised in L2 pronunciation instruction and assessment? (p267-8). McNamara explicitly states why understanding the presence of features on a rating scale is such an important consideration by arguing that:

“we must remain sceptical about the meaning of our test scores, and do everything we can to improve our understanding of what they mean, in the interests primarily of fairness to the test candidates”

(McNamara, 1996: 246)

Therefore, the attempts of this study to overcome the challenges associated with developing pronunciation descriptors are timely and appropriate. This study is significant in presenting a perspective on the possibility of increasing the level of detail on pronunciation criteria. In other words, bridging the gap between what features researchers identify as objectively associated with scores by establishing the degree to which test takers actually attend to them consistently. Furthermore, the presentation of rating scale descriptors derived from the data collected for this study provides considerable support and guidance for such features on rating scales.

1.4 Important Terms Defined

There are several terms used in this thesis which have specific and technical meanings. Although some of these terms are discussed in detail in Chapter 2, for the avoidance of confusion important terms are briefly defined here.

Intelligibility

Intelligibility is notorious for being defined in a broad range of different, and sometimes contradictory, ways. In this thesis the term is defined, following Field (2005: 401), as listener perception of recognition of the phonetic content of the speaker's speech signal. This is often measured by observing how effectively a listener can transcribe a speaker's delivery. For consistency with the context of this study, and to limit the methodological constraints associated with transcription of long extemporaneous speech, it is measured by scalar judgement in this thesis. The term contrasts with comprehensibility, which has been defined as a listener's judgement of how easy it is to understand a speaker (Derwing, Munro, & Wiebe, 1998: 396). These definitions are examined in more detail below in Section 2.2.

Typical Listener

Listeners have a wide range of proficiency in listening due to a range of factors, such as accent familiarity and linguistic experience. There is no such thing as an average listener. However, for the purpose of this thesis, it is necessary to contrast experienced raters with people who are not experienced, and to describe ways of generalising the interpretations of these specialist experienced listeners to people who do not have such experience. This imaginary average listener is described by Tench as ‘the-man-in-the-street’ (1997). In this thesis, the term *typical listener* refers to a UK-based native-English speaker with limited exposure to foreign accented speech. Similarly, the term ‘real world’ in this thesis refers to the typical communicative encounters that those listeners take part in.

1.5 Outline of the Thesis

This thesis is organised into seven chapters. Here the structure and content of each chapter is summarised.

Chapter 1 – Introduction

This chapter is concerned with describing the motivation for this study, summarising the current status of pronunciation assessment, setting the research context, and highlighting the significance of the study. The issue of suprasegmental features of speech being somewhat neglected in the current research is raised, and researchers attempting to establish the role of such features is introduced. It is proposed that, despite a broad range of studies in this area, there is still more to discover about the way raters interpret both suprasegmental features of speech and the intelligibility construct. The thesis is contextualised by discussing the setting of this research, presenting case studies that illustrate types of speaking tests, and introducing some of the materials used in the study. Following this, the significance of the study is discussed, describing

the implications of this research and pre-empting the construction of a series of suprasegmental rating scale descriptors. Finally, several of the most important terms used in this thesis are defined.

Chapter 2 – Literature Review

The second chapter develops the issues raised in the introduction. Initially, fundamental elements of the study are discussed in detail. This includes a description of suprasegmental features, and a discussion of why stress, rhythm and intonation were selected as the suprasegmental categories that would be examined in this thesis. Intelligibility is also discussed. A rationale is provided for selecting intelligibility as the criterion of interest in this study. The current status of suprasegmental features in reference to intelligibility is then described. This is primarily in reference to instrumental analysis of the speech signal. Finally, the studies which relate rater awareness of such features to their scoring decisions are reviewed. The outcome of this section is the argument that our current understanding of which features are influential to pronunciation scores is not clearly defined, and the way in which raters attend to such features is not well established, especially in reference to the intelligibility construct. The elements that make up pronunciation assessment, frameworks, scales, and raters, are then discussed. Specifically, current frameworks and scales are reviewed, the role of rater experience in judgements is defined, and existing methods of rating scale construction are reviewed. The chapter closes by presenting the research questions and hypotheses which are provoked by the exploration of the literature. Finally, there is a description of the original contribution to the literature that this study makes.

Chapter 3 – Research Design

Chapter 3 describes the design employed to answer the research questions listed in Chapter 2. The chapter describes a mixed-methods design which employs a rating scale to collect scores and a stimulated recall procedure to collect verbal reports. These methods are rationalised, as

are the approaches to analysing and integrating the resulting data. This section refers to many-facet Rasch measurement analysis and qualitative analysis, as well as a strategy for integrating both types of data in order to develop rating scale descriptors. The criteria for selecting participants and the recruitment strategy is explained. The materials that were used for data collection are introduced and the data collection procedure is described in detail. The pilot, which was a small-scale version of the main study, is described and its influence on the methodological approach taken in the main study is discussed. Finally, the ethical considerations of the study are explored.

Chapter 4 – Quantitative Findings and Discussion

Chapters 4, 5, and 6 follow similar patterns. The method of analysis is described, the findings are presented and then there is a discussion. In Chapter 4, the quantitative analysis describes the way item response theory is employed, and the findings focus on how the analysis answers Research Questions 1 and 2. Specifically, it discusses scoring patterns and rater behaviour of both individuals and rating groups. The discussion revolves around highlighting the level of scoring consistency within the experienced rater group and the difference in scoring patterns between experienced and non-experienced raters. Empirical evidence is presented indicating that raters are broadly consistent with one another, and that there are notable consistencies between experience groups.

Chapter 5 – Qualitative Findings and Discussion

The fifth chapter describes the way qualitative analysis techniques were employed to answer Research Question 3. The method of transcription and coding are described, and the way categories were collated is discussed in detail. The importance of matrix queries as an analytical tool is also explained. The findings are discussed in reference to raters and the original analysis of test takers' deliveries. The discussion focuses on the roles of rhythm, intonation, and lexical

stress in raters' judgements. The primary findings of this section are that raters are broadly consistent in the way they attend to intonation and rhythm, although there are certain specific features which they do not appear to find influential. Chapters 4 and 5 form the empirical basis for the construction of a series of rating scales descriptors which reflect rater perceptions of influential suprasegmental features at varying intelligibility levels.

Chapter 6 – Towards Suprasegmental Rating Scale Descriptors

The approach to deriving the scales, specifically the integrated display employed to integrate qualitative and quantitative data, is described in Chapter 6. A 6-point scale containing descriptors for rhythm, intonation and lexical stress is then presented and finally discussed. The discussion takes the form of a comparison of the suprasegmental descriptors with those contained in other rating scales and frameworks. The quality of the resulting descriptors is also examined in reference to established standards, and trialling of the descriptors is recommended as a suitable next step to ensure the usability of the scale.

Chapter 7 – Conclusion

Chapter 7 concludes the thesis by summarising the research approach and methodology, before presenting the key findings. A brief general discussion brings all the strands of this project together by stating how the research questions were resolved and presenting implications of the research generally. A series of potential applications of the descriptors developed in Chapter 6 are explored, before the limitations to the study are described, including reference to the status of native speaking raters and the task types used as stimuli data. Following this, a series of potential future directions for this research are described, including expanding the CEFR levels targeted, as well as manipulating characteristics of the test takers and raters in order to further explore the generalisability of the findings. Finally, a suitable approach to trialling the descriptors is proposed

and there are a few concluding remarks reiterating the unique contribution that this study makes to the field.

Chapter 2: Literature Review

The purpose of this chapter is to review what is currently known about the place of suprasegmental features of speech in the assessment of pronunciation. The first two sections are primarily concerned with defining terms: Section 2.1 describes suprasegmental features, and Section 2.2 discusses *intelligibility* and explains its suitability as a pronunciation criterion. Intelligibility and suprasegmental features are then brought together in Section 2.3, which discusses the current state of research into the link between use of suprasegmental features and listener perceptions of proficiency in pronunciation. Researchers employ two primary approaches to this, the first relies on instrumental measurement of the speech signal and correlates it to proficiency metrics, such as scalar judgements or transcription tasks. The second approach uses verbal reports to collate the features of speech that raters attend to when making judgements. Section 2.4 then introduces rating scale development and design. In this section rating scales are described and current frameworks and scales are reviewed in terms of their attention to suprasegmental features of speech. In Section 2.5, the role of the rater is discussed, particularly in reference to how rater variation influences judgements. The chapter closes by summarising the gaps in the current literature, describing the research questions which constituted the focus of this study, presenting hypotheses, and stating the original contribution to the literature made by this study.

2.1 Suprasegmental Features of Speech

The terms *suprasegmental* and *prosody* are often used interchangeably although they can have different meanings. *Prosody* has roots in classical metrical composition and has been used specifically to refer to pitch, loudness, tempo, and rhythm (Crystal, 2011: 393). By contrast, *suprasegmental* is used by phonologists as a superordinate label for features that accompany consonants or vowels and extend over more than one phoneme.

A definitive list of suprasegmental features does not exist, but the following table gives some indication of the range of speech features that can be characterised as suprasegmental.

Table 2.1: Lists of Suprasegmental Features

Fox, 2000	Laver, 1994	Lehiste, 1970	Crystal, 1969	Crystal & Quirk, 1964	Sweet, 1906
length	pitch	quantity	pitch direction	tone	length
accent (stress)	loudness	tonal features	pitch range	tempo	stress
tone	rate (tempo)	stress	pause	prominence	pitch
intonation	pauses		loudness	pitch range	voice-quality
	rhythm		tempo	rhythmicality	
			rhythmicality	tension	
			tension	quality	
				qualification	
				pause	
				vocalisation	

Adapted from Fox (2000: 10)

For this study the focus rests on stress, intonation, and rhythm. This is for consistency with the context. These characteristics of the speech signal are typically included in rating scales (see Section 2.4 below) and language learning textbooks. Additionally, stress, intonation, and rhythm could be considered superordinate to the range of other suprasegmental features listed in the table above. Length, pitch, and quality contribute to stress; continuity, tempo, and pausing are qualities associated with rhythm; and pitch, tone, and stress are connected to intonation. Finally, although these terms have technical meanings they are also accessible to non-experts and can be measured perceptually. These three suprasegmental terms are now discussed in detail.

2.1.1 Stress

Stress is the prominence of one syllable relative to the surrounding syllables. It can be relative salience of a syllable within a word, known as lexical stress or word stress. Equally it can be relative salience of a syllable within an utterance, referred to as focal stress in this thesis, but also known as sentence stress, intonational stress, nuclear stress, and tonic stress. Focal stress is closely related to intonation and it is difficult to discuss intonation without reference to focal stress.

The marking of syllable prominence is created suprasegmentally in English by a combination of moderating pitch, loudness, quality, and/or length of syllables (Laver, 1994: 512). The relative importance of these features in determining stress patterns is disputed. Full quality vowels in stressed syllables have longer duration (Laver, 1994: 448). Whether that means length is the main constituent of stress is open to question, indeed Fry (1958), Bolinger (1958) and Beckman (1986) argued that pitch is the most reliable indicator of prominence in English. Fry (1958; 1955) places pitch as the most influential feature to stress, followed by duration and, least influentially, loudness. By contrast, in a study employing seven English dialects of the British Isles, Kochanski et al. (2005) found that stress was marked by duration and loudness more than pitch. More

recently Zhang and Francis (2010) suggested that English and Mandarin listeners consider vowel quality (for example, reduction) to be a greater cue to lexical stress than measures of pitch, duration, and intensity, and Cutler (2015: 120) concurs that stress is primarily marked by the pattern of strong and weak vowels in a word.

This variation in understanding of what constitutes stress presents a challenge to teachers and testers in that it clouds precisely what should be taught and assessed. Experienced listeners of English are well able to identify and count stress without the need for explicit instruction (Hammond, 1999: 152), indicating that irrespective of the relative influence of each phonetic element, expert listeners with the same L1 have a common perception of stress.

Lexical stress

English is a lexical-stress language, meaning that multi-syllable words can have up to three types of stress: 1) primary stress, which marks the most salient syllable whose onset is indicated in phonetic transcription by a raised vertical line ('); 2) secondary stress, which is less salient than the primary stressed syllable and is indicated with a lowered vertical line (,); and 3) unstressed syllables, which are not stressed relative to the primary and secondary stresses. For example, *photographic* has primary stress on the third syllable and secondary stress on the first syllable, whereas *accommodation* has primary stress on the fourth syllable and secondary stress on the second syllable:

,photo'graphic /,fəʊtə'græfɪk/

a,ccommo'dation /ə,kɒmə'deɪʃən/

Exceptions are common to the rules for the placement of lexical stress in English; in nouns with two syllables, stress is most likely to fall on the initial syllable although this is not always the case (Cutler & Carter, 1987; Zhang, Nissen, & Francis, 2008). However, lexical stress may be critical to recognition of words, in the sense that appropriate placement of stress may help a listener to select a word against competing alternatives (Cutler & Pasveer, 2006). This has implications for the assessment of speaking because the inclusion of lexical stress on rating scales should reflect its impact on intelligibility.

Focal stress

In English focal stress typically falls on the primary lexical stress of the last content word of an utterance. The purpose of focal stress is primarily to signal the information structure of the utterance. By shifting the focal stress, a speaker can organise utterances, highlighting the most important information for the listener and communicating what information is missing and what is known:

Where did James go on holiday? [neutral]

Where did James go on holiday? [James rather than anyone else]

Where did James go on holiday? [which location specifically]

In the above example, focal stress is a critical tool to communicate meaning, and this is expanded upon in the following section. Although lexical and focal stress have some parallels in the way they are articulated, the function of focal stress is closely linked to that of intonation. Therefore, from now on focal stress will be dealt with as a feature of intonation.

2.1.2 Intonation

Intonation Form

Intonation is the change in pitch that occurs over the course of an utterance, where pitch is a function of the rate of vibration of a speaker's vocal folds. In tone-based languages, such as Mandarin Chinese, pitch can have a lexical function. This is not the case in English, which allows pitch to be used independently of lexical meaning.

Both the location and direction of pitch movement are relevant to meaning. Laver (1994, citing Halliday, 1963; 1967; 1970) suggests limiting analysis to the following five pitch patterns:

Name of tone	Pitch pattern	Contour name
Tone 1	\	fall
Tone 2	/ \	rise or fall-rise
Tone 3	—/	rise
Tone 4	/ \ /	rise-fall-rise
Tone 5	\ / \	fall-rise-fall

(Laver, 1994: 495)

Several researchers have attempted to assign a meaning to pitch movements. A difficulty is that pitch is relative, in the sense that pitch level or movement is meaningful only in relation to listener perception of adjacent pitch levels. This means that its interpretation depends on a complex interaction between production and perception. This is fraught with difficulty, not least because there is intonation variation among varieties of English. For example, Grabe et al. (2005) recorded utterances from a range of locations in the British Isles and transcribed the intonation of each

speaker. They found broad variation in intonation patterns used for questions and statements and argued that “learners of English need to be prepared for extensive variation in the intonation they might hear from native speakers” (p21). In other words, even native speakers of English do not necessarily deliver intonation in a consistent fashion. This has direct implications for the construction of pronunciation descriptors that can be applied reliably.

Intonation Function

The functions of intonation can be described in relation to the domains of discourse, grammar, accent, and attitude. The *discourse* function of intonation describes the way intonation can be used to link or contrast material within a discourse, or to signal to the listener what type of response is expected (Sacks, Schegloff, & Jefferson, 1974). It is linked to turn taking in the sense that the speaker’s intonation contour can provide cues as to whether they are prepared to cede their turn to another speaker. According to Roach (2000: 153-4), falling pitch denotes finality, and rising pitch suggests there is more of the message to come, or invites the interlocutor to continue.

Discourse is also managed by using pitch to signal new versus given information. Low pitch accent is linked to information which is already available to the listener and requires minimal attentional effort (Chafe, 1994: 169), such as certain discourse markers (e.g. “you know”) (Wennerstrom, 2001: 37). This is closely linked to the *accentual* function of intonation, where relative salience is manipulated to apply emphasis. In neutral utterances, focal stress typically falls on the lexical stress of the final content word in the utterance, as in 1a below. A speaker can manipulate this to create emphasis, as in 1b:

- 1a. I was very happy [neutral placement]
 1b. I was very happy [adds additional emphasis]

It can also be used to demonstrate contrast. For instance, in 2a below the speaker contrasts *Italy* with other locations, and in 2b the speaker contrasts *this* year with other years. A steep rising pitch is likely to be assigned to the salient syllable in these utterances (Wennerstrom, 2001: 36):

2a. I'm going to Italy this year [not any other country]

2b. I'm going to Italy this year [not any other year]

An intonation contour can operate *grammatically* by signalling syntactic boundaries. It helps the listener to follow the meaning of an ongoing utterance. This is exemplified below in 3a and 3b, where the same utterance is split into two tone units at different points. A tone unit is part of an utterance that contains one tone-modified syllable (known as a tonic syllable), which is stressed (known as tonic stress, or nucleus). In 3a and 3b, the tone unit boundary is placed before or after the word *quickly* by the speaker. Phonetically, the speaker does this by varying the pitch level on the first and last syllable of each tone unit and the pausing between them.

3a Those who spoke quickly | got an angry response

3b Those who spoke | quickly got an angry response

(Ashby & Maidment, 2005: 167)

The effect is to fundamentally change the meaning of the utterances. In 3a the adverb *quickly* modifies the verb *spoke* whereas in 3b it modifies *got*. There is debate over the link between grammar and intonation; in an article entitled *Accent is predictable (if you're a mind-reader)*, Bolinger (1972) argues that intonation cannot be neatly linked to grammatical function because other functions of intonation can interfere. In addition, Nolan (2006: 441) states that intonation is

not always clearly marked in fast natural speech, which raises questions as to its criticality for effective communication.

Finally, certain aspects of a speaker's *attitude* may be discernible from intonation (Cowie et al., 2000; Cowie, Douglas-Cowie, & Romano, 1999). According to Roach (2000: 153-4,57), a level pitch may indicate that the speaker is bored, a fall-rise could show limited agreement, especially in British English (Levis & Wichmann, 2015: 149), and a rise-fall approval, disapproval or surprise. Pickering (2001) argued that overuse of falling tones by teaching assistants resulted in a perception of "disinterest or lack of involvement" (p251). However, there are few consistent phonemic markers of attitude and there are a range of emotions that cannot be consistently identified on the basis of the speaker's pitch contour (Pittam & Scherer, 1993, cited in Levis & Wichmann, 2015: 148).

In summary, intonation has a complex relationship to meaning. In 2005 Vaissière argued "there is as yet no comprehensive model of intonation perception which includes the interaction between the various often conflicting functions of intonation" (2005: 237). This is still the case, and precisely how intonation operates in the relatively constricted context of the language test is not well understood, as discussed below in Section 2.3.

2.1.3 Rhythm

Speech rhythm is a listener's perception of a regular pattern of events in the speech signal. It has been known for some time that speech is delivered in regular beats (Classe, 1939; Pike, 1945). The function of rhythm in English appears to be to facilitate lexical segmentation, although the relationship between rhythm and segmentation is not simple one:

“its role is not to provide direct signals of word boundary location, but rather to establish a framework within which listeners can orient their hypotheses as to most probable word-boundary locations”

(Cutler, Dahan, & Van Donselaar, 1997: 148)

In English, rhythm is typically regarded as the pattern of strong and weak syllables. However, there has been extensive debate as to the role of different types of rhythm timing. Abercrombie (1967) studied the difference between speech with equally timed syllables and speech with equally timed stresses. He suggested that languages have rhythmic timing derived from either regularity in delivery of syllables (syllable-timing) or regularity in delivery of stresses (stress-timing) (1967: 97).

Laver (1994) regards the view that rhythm can be timed by stress or syllable weight as “tenacious but controversial” (p156). Indeed, perceptual studies are contradictory, demonstrating in some cases that listeners can distinguish between languages classed as stress- or syllable-timed (Ramus & Mehler, 1999) even from an early age (Nazzi, Bertoncini, & Mehler, 1998), and in other cases that they cannot (Scott, Isard, & de Boysson-Bardies, 1985). In fact, many early studies do not corroborate Abercrombie’s speech timing hypothesis (e.g. O’Connor, 1965; Hill, Jassem, & Witten, 1979; Faure, Hirst, & Chafcouloff, 1980; Dauer, 1983). More recently the development of instrumental tools for measuring speech rhythm, such as the Pairwise Variability Index, which measures the average difference between successive vocalic and consonantal intervals, have been used to identify differences between syllable- and stress-timed languages (Grabe & Low, 2002). These metrics are widely used to investigate L2 learners (e.g. Díez, Dellwo, Gavaldà, & Rosen, 2008, Ordin, Polyanskaya, & Ulbrich, 2011), including by Galaczi et al. (2016) who demonstrated that L2 English speech becomes more stress-timed as learners progress to higher proficiency levels. However, there is debate over whether rhythm metrics are capable of

measuring timing. They have been found to lack robustness (Arvaniti, 2012; Arvaniti, 2009; Wiget et al., 2010). For instance, Arvaniti and Rodriguez (2013) modified recordings of a series of languages to retain the timing but remove other characteristics. Listeners were then asked whether a series of languages belonged to the same group as English. The researchers found that listeners did not appear to differentiate between languages on the basis of timing alone.

Clearly the status of timing is not well established. Current thinking falls into two strands. The first retains that stress- and syllable-timing are instrumentally measurable but hypothesises that they are on a continuum with languages tending towards one end or the other (Grabe & Low, 2002). The second strand suggests that the phenomenon of speech rhythm is too perceptually subjective to be captured using phonetic techniques (Gut, 2012). Laver (1994: 524) suggests that the human cognitive system is drawn to applying rhythm to speech and other events in time. He cites evidence showing that listeners respond more rhythmically to speech than is justified by the material they are exposed to (Allen, 1975; Beckman, 1992), indicating that rhythm may indeed be partially imposed by the listener rather than the speaker.

The position taken in this thesis is that rhythm is an abstract term applied to a constellation of perceptual phenomena. It is possible to measure certain of these phenomena instrumentally, but the link between measurable phonetic reality of rhythm and perceptual realisation of rhythm is not clear. Analysis of rhythm-related characteristics of the speech signal can only provide limited practical insight into the way human scorers interpret rhythm in language assessments. One aim of this research, therefore, is to identify whether this interpretation is consistent from listener to listener, to such a degree that it can be assessed reliably in a speaking test.

The suprasegmental categories used in this study have been defined and described, and attention now turns to intelligibility.

2.2 Defining Intelligibility

The term *intelligibility* is widely used in pronunciation research but the way it is defined and measured varies greatly. In broad terms *comfortable intelligibility* is described as “a pronunciation which can be understood with little or no conscious effort on the part of the listener” (Abercrombie, 1949: 120). This is a useful general introduction to the term but it suffers from a lack of precision, which would make it unsuited to analytical approaches to rating scale development. A narrower definition of intelligibility was provided by Smith and Rafiqzad (1979), who regarded intelligibility as word recognition. This definition was taken up by Smith and Nelson (1985), who called for use of the terms *intelligibility*, *comprehensibility*, and *interpretability* to be standardised. They argue for *intelligibility* to be defined as a listener’s word or utterance recognition of the phonological elements of the speaker’s speech signal. They describe *comprehensibility* as the ability to grasp word or utterance meaning, the locutionary force, and interpretability as speaker intention, the illocutionary force. This definition provides a psycholinguistically relevant separation between *comprehensibility* as global processing and *intelligibility* as local recognition (Field, 2005), and in this formulation intelligibility operates as a gatekeeper of effective communication (Jenkins, 2000). Although intelligible speech is not necessarily comprehensible, comprehensible speech must be intelligible.

Munro and Derwing (2015a: 378) argue that regarding intelligibility as a component part of comprehensibility may make it impossible to clearly measure perceptual recognition separately from understanding. This is due to listeners using their global understanding of an utterance to infer the form of otherwise unrecognisable words. On these grounds Munro and Derwing prefer a definition of *comprehensibility* to be listener perception of ease of understanding, often measured by subjective judgements on a rating scale and *intelligibility* to be actual understanding, typically measured objectively using tasks such as transcription. In effect, this disrupts the separation

between recognition and understanding implied by Smith and Rafiqzad (1979) and Smith and Nelson (1985), because it includes reference to understanding in the definition of intelligibility. In other words, Derwing and Munro are concerned with the separation between objectively and subjectively measured understanding, whereas Smith and Nelson focus on the difference between recognition and understanding.

Intelligibility, in Smith and Nelson's terms, therefore, is predominantly influenced by a speaker's ability to articulate recognisable utterances, and the listener's ability to interpret the phonetic cues within those utterances. This contrasts with comprehensibility, which depends initially on recognition of the speech signal, but also involves linguistic elements besides pronunciation, such as lexis choice, syntax, discourse logic and pragmatics. A natural starting point in the assessment of pronunciation effectiveness, therefore, is the measurement of intelligibility. This is because intelligibility is distinct from locutionary and illocutionary force and is therefore influenced less by factors that employ higher processing to develop understanding.

Intelligibility, whether characterised as perceptual recognition or actual understanding, is typically measured using transcription tasks (e.g. Derwing & Munro, 1997; Munro, Derwing, & Morton, 2006). It is also sometimes measured using comprehension questions (Smith & Bisazza, 1982) and cloze items (Smith, 1992; Browne & Fulcher, 2016). These approaches conflate information based on context and co-text with information that is perceptually derived. Practical limitations on the transcription method for the assessment context are that objective tasks may not be able to distinguish gradients of intelligibility or reflect the different levels of processing required to recognise different qualities of non-native speech. The blunt nature of transcription can be demonstrated by Derwing and Munro's (1997) study in which 60% of transcriptions were error free, suggesting these speakers were totally intelligible. Furthermore, transcriptions do not account for rater interpretation. In human-scored assessments, it is the rater's perception of the

speech signal that is the basis of scores, rather than any objectively measured metric designed to reflect intelligibility or understanding.

In the current study intelligibility is defined, in line with Smith and Nelson (1985) and following Field (2005: 401), as listener perception of recognition of the phonetic content of the speaker's speech signal. Henceforth, the term *intelligibility* refers to this definition of intelligibility as local perceptual recognition. *Comprehensibility* is reserved for references to global understanding.

2.3 Suprasegmental Correlates of Intelligibility

A great deal of research has been undertaken to identify the influence of suprasegmental features on pronunciation constructs. Hahn (2004) traces the link between suprasegmental features of speech and understanding back to Nida (1957), who argued that speech containing appropriately expressed individual sounds and grammatical forms could still be difficult to understand due to unnatural intonation. Indeed, suprasegmental errors are often characterised as more damaging to oral proficiency than segmental errors (e.g. McNerney & Mendelsohn, 1992; Caspers, 2010: 17; Anderson-Hsieh, Johnson, & Koehler, 1992). Limited empirical support for this assertion is what prompted Munro and Derwing's (Derwing & Munro, 2005:386; Munro & Derwing, 1999) research into intelligibility in foreign-accented speech. They recognise that there are a number of studies identifying hierarchies of pronunciation errors (e.g. Albrechtsen, Henriksen, & Faerch, 1980; Johansson, 1978; Schairer, 1992) but argued that the variety of research contexts makes generalisability difficult (Munro & Derwing, 1999:288).

Research in this area follows two main routes. In some cases, researchers instrumentally measure the speech signal and correlate the findings with score or other proficiency measures, such as transcriptions. This is reviewed below in Sections 2.3.1 – 2.3.3 covering lexical stress,

intonation and rhythm. In other cases, researchers use qualitative approaches to examine the features that raters attended to when making decision. This is discussed in Section 2.3.4.

2.3.1 Lexical Stress

There is contradictory evidence for the role of lexical stress in recognition of speech. Many studies find misplaced lexical stress to be related to reduced intelligibility (Anderson-Hsieh et al., 1992; Bond, 2008; Bond & Small, 1983; Van Donselaar, Koster, & Cutler, 2005; Zielinski, 2008), while others find that stress is only influential insofar as it is related to vowel quality (Cooper, Cutler, & Wales, 2002; Cutler, 1986). A recent study by Jesse et al. (2017) employing eye tracking software reported on an experiment that consisted of playing a word to listeners and asking them to select it from a series of onscreen options. The options were segmentally identical for the first two syllables but suprasegmentally different (e.g. *admiral* versus *admiration*). They found that until the words became distinguishable in segmental terms (i.e. by the third syllable), listeners spent more time looking at the targeted word if it had stress on the first syllable, indicating that listeners were using lexical stress to distinguish words from competitors.

Transcription tasks are a common way of identifying the role of lexical stress on recognition. Field (2005) manipulated lexical stress and vowel quality in a series of words and played them to native and non-native English speaking listeners. He found that shifting lexical stress influenced the listeners' ability to accurately transcribe words, irrespective of the role of vowel quality. This was an expansion of the study by Cutler and Clifton (1984), who asked subjects to listen to a series of correctly and incorrectly articulated words and to decide whether they were correct or not. They found that the combination of vowel quality and stress change had a greater impact on speed of word recognition than stress changes alone. Certain laboratory studies, however, suggest correct lexical stress may be less closely linked to intelligibility in the case of noise-masked words

(Slowiaczek, 1990) and minimal pairs discrimination (Cutler, 1986; Small, Simon, & Goldberg, 1988). The balance of evidence, therefore, seems to indicate that there is a role for lexical stress in intelligibility in some contexts and using certain methods of enquiry.

In the second language assessment context, there is limited research into the role of lexical stress in intelligibility. Isaacs and Trofimovich (2012) examined the role of lexical stress in rater judgements of comprehensibility, defined as ease of understanding. They instrumentally measured the suprasegmental features of 40 samples of French learners of English and asked native English judges to listen to the samples and rate them on a 9-point scale for ease of understanding. The Pearson correlation coefficient indicated a strong relationship between judgements of ease of understanding and lexical stress error ratio ($r = -0.76$, $p < 0.01$, two tailed), and vowel reduction ratio ($r = 0.74$). Coupled with the findings of Field's transcription tasks and Cutler and Clifton's binary choice task, it seems to indicate that lexical stress has an impact on judgements of proficiency in a language testing context, even if the precise impact on recognition is not well established.

2.3.2 Intonation

Vaissière (2005) regards intonation as difficult to study because there is no universal definition of it, no consensus on the purpose of research into it, and no agreement as to how it should be represented. It is unsurprising then that research findings are mixed. Nonetheless, intonation appears to influence scores on a range of pronunciation proficiency criteria. For instance, Kang (2013b) reported that pitch accompanied stress as a contributor to listeners' impressions of overall oral proficiency, but identifying the details of precisely which aspects of intonation are influential presents a somewhat more complex picture. Wennerstrom (1998) measured pitch variation in speech samples of 18 Chinese L1 speakers delivering lectures in English and then conducted a

multiple regression analysis of pitch variation and the scores they received on a global comprehensibility test. She found the use of broader pitch range at rhetorical junctures to be a significant predictor of global comprehensibility. Notably though, she also found that pitch used in several other ways, such as in content words versus function words, and at phrase boundaries, was not significantly related to test scores. This raises the possibility that the use of pitch in some contexts might be more influential to pronunciation proficiency than in others, but the study is somewhat limited in the range of features it examines.

Further insight is provided by Pickering (2001). She recorded lectures by native and non-native teaching assistants and measured the number of rising, falling and level tones used, as well as the functions of these tones in the discourse. She found some significant differences between the way native-speaking and non-native-speaking teachers used tone. Native speakers used more rising pitch, which Pickering (following Brazil, 1997) regarded as helping the speaker build rapport with students and make the lecture more accessible to them. The lack of rising tone in non-native speaking teachers and the corresponding increase in flat and falling tones, by contrast, results in the deliveries appearing to “create a flat, monotonic pitch structure unfamiliar to NS [native-speaking] hearers” (Pickering, 2001: 249). Critically though, Pickering does not measure the impact on listeners and therefore it is not clear whether the interpretations of errant tone applied by Pickering translated to reduced intelligibility, or indeed any communicative problems.

In the second language testing context, Kang et al. (2010) expanded on Pickering’s findings by measuring both pitch in the speech signal and the impact on listeners, as well as by using stimuli material taken from an authentic examination, the TOEFL iBT. They measured nine different tone choices (e.g. high-rising, low-level) in addition to pitch range and pitch on prominent and non-prominent syllables, and on new versus given information. They related those measures to judgements of overall oral proficiency and comprehensibility using multiple regression analysis.

They found that the best predictor of oral proficiency was a cluster of features described as “suprasegmental fluency” in addition to high-rising and mid-rising pitch. The suprasegmental fluency cluster of features included various pace measures in addition to stress and mid-falling tones. Overall, the suprasegmental features that they measured were found to account for 50% of the variance in scores for oral proficiency ($R^2 = 0.51$) and comprehensibility ($R^2 = 0.50$). The study is ground breaking in its scope of suprasegmental measures and the range of oral proficiency and comprehensibility ratings. It effectively fills in the gaps left by Wennerstrom, in terms of precisely which elements of intonation are influential, and by Pickering, in terms of the impact on rater judgements. Furthermore, it lays the groundwork for other studies which have provided additional insight into the links between oral proficiency and intonation. For instance, Isaacs & Trofimovich (2012) found a moderate correlation between comprehensibility judgements and pitch contour ($r = 0.57$, $p < 0.01$, two tailed), although interestingly, pitch range was not found to be significant in the same study, and Saito, Trofimovich, & Isaacs (2016) found that that the magnitude of the influence of intonation may depend on speaker proficiency level.

The study by Kang et al. (2010) is certainly wide-ranging, however, three areas relevant to assessing pronunciation remain unexamined: 1) the study makes no reference to the participants’ interpretations of the speech signal. Native speakers with different accents and dialects can apply varying intonation contours for the same function, and the acoustic reality measured by Kang et al. is not necessarily a complete reflection of the psychological reality of listener perception (Shoemaker & Rast, 2013: 166), meaning it is difficult to see how these results could be operationalised on a rating scale for use by human raters; 2) the use of undergraduate judges with no experience in rating language means the resulting scores may not reflect the results of an assessment encounter; and 3) neither the oral proficiency nor the comprehensibility measures focus solely on pronunciation. The oral proficiency scores were derived by summing the scores on six 7-point Likert items, which include pronunciation and accent, grammatical accuracy,

vocabulary, rate of speech, organisation, and task completion. The comprehensibility measure similarly combined judgements of understanding, comprehensibility, effort, clarity and difficulty. Although the comprehensibility criterion appears to measure overall oral proficiency, it neglects intelligibility.

Despite these gaps there is good evidence that broad interpretations of intonation have an impact on judgement of global proficiency. Iwashita et al. (2008) assessed how many complete tone units test takers produced and then how English-like the intonation was by placing them in one of three categories (English-like, nearly English-like, and non-English-like) using a range of criteria related to pitch movement and L1 interference. The researchers found that speakers who received a high score on a 5-point global proficiency scale were regarded as having more tone units and for those being more English-like. This is a clear indication that global control over intonation is related to pronunciation proficiency. There is less evidence in reference to intelligibility.

Intelligibility

Two important themes of research into the effects of L2 intonation will now be discussed, the influence on intelligibility, and the role of focal stress. It was argued in Section 2.2 above that intelligibility is the most apt criterion against which to assess pronunciation. Evidence for the influence of intonation on intelligibility is mixed. Whereas speech with an artificially flattened pitch profile has been found to be less intelligible, as measured by a transcription task, than natural speech (Laures & Weismer, 1999), two auditory studies by Derwing and Munro (Munro & Derwing, 1995a; Derwing & Munro, 1997) contrast with this. In the 1995 study they asked 18 native English listeners to listen to audio clips of Mandarin speakers of English, transcribe what they heard, and rate it for accentedness and comprehensibility on a 9-point scale. The transcriptions were scored for accuracy of transcription to arrive at an intelligibility score. The

researchers calculated the correlation of each of these criteria against their own judgements of 'nativeness of intonation', and found that although intonation was linked to accentedness and comprehensibility, it was not well correlated to intelligibility. A similar approach was taken in their 1997 study where researcher judgments of 'prosodic goodness' were not found to be related to intelligibility for most listeners. This may be a methodological issue. Intelligibility was found to be very high in both the studies as "more than half of the transcriptions received intelligibility scores of 100% and many others contained only minor errors" (Munro & Derwing, 1995a: 89). Such a high score is likely to confound calculations of correlation.

Focal stress

Focal stress has often been examined separately from intonation but the findings relating it to intonation are no more clear cut. This is primarily due to the range of methods employed. Hahn (2004) recorded a speaker three times, once with appropriately placed focal stress, once with inappropriately placed focal stress, and once with no focal stress. Using written recall and a comprehension quiz, she discovered that when the speaker delivered a lecture with appropriately placed focal stress, listeners remembered more main ideas and more information in total. Furthermore, in response to a series of Likert items, listeners who heard the version with appropriately placed stress regarded the speaker more positively. Interestingly though, utilising a dual-task paradigm where listeners had to listen for a tone while also trying to listen to the lecture, she found that, although listeners who heard the speaker with correctly placed stress responded to the tone more quickly, they did not do so significantly according to a one-factor analysis of variance ($p = 0.19$).

These findings were interpreted to indicate that errant placement of focal stress did not have a significant impact on the processing of speech, but clearly there is some impact since it appears

to influence both retention of information and subjective judgement of speaker quality. The finding that listeners regard a speaker more positively if they place focal stress appropriately is supported by Birch and Clifton (1995) who asked listeners to judge prosodic appropriateness of stress-modified sentences on a 5-point Likert item. They found that appropriate placement of stress (on new versus given information) resulted in higher acceptability judgements.

A similar pattern emerges in studies where overall proficiency scores are collected. Kang (2013b) found a relationship between Cambridge ESOL proficiency levels, as assigned by Cambridge Assessment raters, and stress. She instrumentally analysed 120 audio recordings to measure indicators of focal stress, such as the proportion of prominent words to total words and the number of prominent words per run. By undertaking a step-wise multiple regression she identified that 70% of the variance in proficiency levels could be attributed to the pronunciation features that she measured, of which 30.9% was from stress and pitch. Kang's analysis conflates lexical stress, focal stress and pitch range, which makes it difficult to identify precisely whether focal stress is an important factor in this study, and it also does not target misplacement of stress but rather metrics of stress use. Similar issues arise in a study by Chen and Wang (2016) who found that neither the number of stressed words per minute nor the proportion of stress words were significant contributors to attractiveness judgements.

The influence of intonation on pronunciation proficiency broadly is well-examined, and several characteristics of intonation, including focal stress, certainly influence scores. Focal stress appears to be relevant to rater judgements of speaker quality, and it also appears to influence the ability to remember information. However, in studies specifically dealing with intelligibility, there is limited evidence and the existing findings are contradictory. In terms of the instrumental analysis presented here, further research is required to establish the nature of the link between intonation and intelligibility.

2.3.3 Rhythm

Several studies have examined the role of rhythm in the context of language testing. Evidence for the importance of rhythm to intelligibility is provided by Zielinski (2008), who asked listeners to transcribe L2 speech then instrumentally examined the parts of the speakers' performances that corresponded to errors in the transcriptions. She found that listeners relied on syllable stress patterns to recognise utterances. In other words, the number and pattern of strong and weak syllables appeared to be related to intelligibility. Raters also appear to regard rhythm as important. In a study by Galaczi et al. (2012), they examined which features of speech raters attended to when making judgements on the International English Language Testing System (IELTS). The findings were that 83% of raters regarded rhythm as salient and 76% regarded stress-timing as salient.

Further insight using rhythm metrics is provided by Galaczi et al. (2016). The researchers measured 20 speech samples for characteristics associated with stress, such as duration between accented and unaccented syllables, and the variability in vocalic and consonantal interval durations. The speech samples were assigned a level on the Common European Framework of Reference for languages (CEFR, discussed in Section 2.4.2 below) according to the pronunciation score they received in a Cambridge English exam. By observing the difference in rhythm measures over different CEFR levels, they could track how rhythm changed as speakers became more proficient. For example, as speakers were judged to have better pronunciation the durations of accented and unaccented syllables tended to change, the number of syllables they produced tended to increase, and the duration of syllables tended to shorten. Notably, between CEFR levels B1 and B2, speakers gained better control over reduced vowels and reduced the number of incorrectly inserted vowels. This suggests that there is something important happening to rhythm at this CEFR level.

Galaczi et al. (2016) make some interpretations of rhythmic timing by examining the difference in rhythm metrics by German, Spanish and Korean speakers of English. They argue that as speakers progress in proficiency they also change from a delivery that is more typically syllable-timed to one more stress-timed. Other studies corroborate this, as shown by Tajima et al. (1997) who manipulated the English speech of Chinese speakers and asked listeners to identify which words were being said. The outcome was that intelligibility improved when native-like timing was applied to non-native speech. These findings are illuminating but the criticism of rhythm metrics in instrumental studies of speech timing must be reiterated. Arvaniti (2009) presents a strong view in this regard, arguing that rhythm metrics are unreliable predictors of rhythm because they focus on timing and durational variability rather than on “groupings and patterns of prominence” (p57). She used rhythm metrics to measure the timing of a series of languages, finding that methodical choices and inter-speaker variation had a greater influence on the results than target language (Arvaniti, 2012).

Iwashita et al. (2008) avoided using controversial rhythm metrics by relying on expert judgements of rhythm timing (as ‘stress-timed’, ‘syllable-timed’, ‘variable’, or ‘unclear’). The researchers compared the assignment of timing to the scores speakers received on a draft global proficiency scale. They found that speakers regarded as low proficiency according to the global judgements tended to have speech that was considered by phoneticians to be ‘unclear’ or ‘syllable-timed’. For the study reported in this thesis, which takes language assessment as a context, this is an important variation from studies that measure rhythm instrumentally, since it is grounded clearly in listener perception rather than acoustic reality. However, it is not clear in Iwashita et al.’s study what criterion rhythm is measured against (e.g. comprehensibility, fluency). Additionally, it still relies on the potentially unhelpful distinction between stress timing and syllable timing. At the very least it provides good information as to the influence of timing on pronunciation proficiency, with the caveat that rhythm is more than just timing.

Rhythm research relies on a preponderance of studies focusing on the role of timing and instrumental measurement of the speech signal. As stated above (Section 2.1.3), the position taken in this thesis is that rhythm is holistic and perceptual, meaning such studies are unable to adequately reflect the experience of raters when making judgements. In addition to this, there is very limited exposition as to the role of rhythm in intelligibility.

2.3.4 Rater Awareness of Suprasegmental Features

The previous section primarily focused on instrumental studies which measured the suprasegmental feature of the speech signal and reported on the influence on the listener. None of those studies asked the listeners directly how they interpreted the speech signal. This section is devoted to identifying the features that listeners regard as influential to their judgements of pronunciation proficiency.

Research into feature attention in speaking tests falls into three categories: 1) studies which make no reference to specific suprasegmental features; 2) studies which mention such features but undertake limited analysis or reporting; and 3) studies which address rater interpretation of suprasegmental features and report them in detail.

In the first category, even when qualitative data is collected, researchers do not report suprasegmental features as influential to scoring decisions. For example, in Orr's study (2002), which consisted of 32 participants reporting how they assessed a paired interview, pronunciation references are all global and do not mention suprasegmental features (p148). Similar findings are described by Kim (2009) who identifies the categories of errors mentioned by participants but not the specific features, and by Winke et al. (2011), and Winke and Gass (2013), who go into no

finer detail than to mention that participants attended to 'quality of voice'. Zhang and Elder (2011; 2014) aimed to identify whether native speaker status is related to the features participants focus on when applying a rating scale. The scale specifically refers to stress and intonation (Zhang & Elder, 2014: 325), yet the researchers only report rater reference to each scale criteria at a high level, as well as categories of non-criterion features. The lack of fine detail makes it unclear if and how raters interpreted and applied stress, intonation and rhythm.

A possible reason for the absence of suprasegmental features in these studies is that the research design meant they did not come under scrutiny, or the research aims meant they were not reported. This is supported by the fact that the aims of some of the studies are not relevant to a detailed discussion of pronunciation characteristics. For example, Winke et al. (2011) were primarily concerned with uncovering bias caused by language background. Although for others, such as Orr's (2002) study using a paired oral task, a discussion of intonation (especially in relation to turn taking) would be wholly appropriate. Furthermore, certain of the studies relied on written rather than verbal reports, such as Zhang & Elder (2011), which is likely to result in much less rich qualitative data. In a study examining how EFL teachers judged learner speech, Gui (2012) noted that "none of the three Chinese raters recorded their opinions about participants' pronunciation in the written comments" (p197), indicating that the method of data collection or the sample size might have influenced the outcome. This is puzzling in some cases, though, since certain of the scales make explicit reference to these features. For instance, in Winke et al.'s (2011) study, participants scored speakers against the TOEFL iBT speaking test which, as discussed below in Section 2.4.3, refers to intonation and rhythm. It is likely that a combination of research aims, methodology, and researcher level of focus explain why suprasegmental features are absent from these studies.

The second category of studies consists of research that provides hints as to the role of suprasegmental features but does not undertake detailed analysis. May (2011) used a stimulated recall procedure to identify what listeners attended to when making judgements. There is reference to 'hesitation' and 'voice quality', which suggest rhythm, and stress may be being considered, as well as the 'extent to which the interaction resembles authentic discussion', which may also encompass suprasegmental features insofar as they are used to manage discourse. Nonetheless, the level of analysis is too high to reveal precisely if and how these features are influencing decisions.

A similar issue arises in Kim (2015), where the author notes that raters with limited experience focused "more on the prosodic features than on the enunciation of each word" (p249), but does not explore specifically the range and magnitude of this behaviour. Equally, Ang-Aw and Goh (2011) collected verbal reports from seven listeners finding that salient factors included 'delivery', 'tone', 'pronunciation', and 'voice'. The researchers considered these outside the scope of the mark scheme and thus did not elaborate on which specific way the participants referenced these. They also provided raters with a questionnaire specifically asking them to report on how important stress, rhythm, and intonation were to their assessments of pronunciation proficiency. However, the outcome of this part of the questionnaire is not elaborated upon in the paper, except to state that "data from the questionnaire showed that raters believed all the features to be important" (p38). Other studies similarly find that raters appear to attend to suprasegmental features, but the researchers do not elaborate on it well (Wei & Llosa, 2015; Ducasse & Brown, 2009).

It is possible that this limited reporting of suprasegmental features is due to participants not being attentive to suprasegmental features. In a study into the features teachers reported attending to when judging comprehensibility, Isaacs and Trofimovich (2012) found that word stress and intonation were considered relatively rarely in contrast to grammar, vocabulary, and fluency. This

is understandable given that comprehensibility requires control over grammar and vocabulary, in addition to effective pronunciation. Here, as with other studies, it appears that researcher aim and methodological approach may be responsible for limited attention to suprasegmental features.

In the final category are studies which investigate pronunciation features in detail. Brown et al. (2005) coded raters' verbal reports for a range of categories including phonology. They found that raters discussed the importance of native-like modulation, the value of using effective information units, and stress placement (p22). At a lower proficiency level, stress, rhythm, and intonation were considered 'extremely nonnative' and 'monotone' (p37). Intonation then became 'rather flat', and misplaced stress and intonation caused progressively less strain as the proficiency level increased. Nativeness was also mentioned by raters in reference to accent judgements in Hayes-Harb and Hacking's (2015) study. Comments such as "native melody" and "stress was off" demonstrate that raters were attending to suprasegmental features of speech. They also reported that raters responded to the speaker's intonation in detail (e.g. "monotone", "correct up preceding commas, down, before periods", p57). Rossiter (2009) also found raters attending to monotonicity and "bad rhythm" (p403) in reference to fluency.

Yates et al. (2011) asked raters to state which features of speech they felt were important when assigning pronunciation scores. The raters stated that global features were the most important, but then in response to a questionnaire, it was revealed that they may in fact be paying more attention to specific suprasegmental features of speech. Raters considered features such as stress, rhythm and intonation to distinguish higher level performances, and during a verbal protocol procedure mentioned intonation more than any other feature. Several other studies with qualitative elements have found that raters discuss stress, rhythm and intonation to some degree (Hubbard, Gilbert, & Pidcock, 2006; Yan, 2014; Isaacs et al., 2015).

These studies make it clear that raters do attend to suprasegmental features of speech when making proficiency judgements, but analytical focus and research design clearly have a large impact on findings. The studies cited above leave several areas unexplored. It is not clear how suprasegmental features influence a rater's decisions specifically in relation to pronunciation, and especially intelligibility. The range of criteria used in existing research demonstrates that suprasegmental features may influence a series of different speaking constructs, and this variety in constructs may explain the variation in features that raters attend to between different studies. However, it also limits the commensurability of these studies, which is exacerbated by the variation in background of speakers and listeners, as well as the speaker proficiency level and speaking context. The existing research is not comprehensive enough to establish how control of suprasegmental features of speech tracks progress in pronunciation proficiency insofar as language raters perceive it.

Summary

Three primary themes emerge from the review of research into suprasegmental correlates of effective pronunciation:

1. Control over stress, intonation, and rhythm appears to be linked to various oral proficiency criteria to varying degrees.
2. Intelligibility, defined as perceptual recognition, is under-researched in this area. Examining the influence of suprasegmental features on intelligibility, measured using a transcription task, is uncommon. It appears not to have been measured using a rating scale in the context of English language assessment before.
3. In studies addressing what listeners attend to when judging the speech signal, there is considerable variation in how participants' interpret suprasegmental features. This raises questions as to their applicability for English language assessment.

There is evidence, both instrumental and perceptual, indicating that listeners do attend to suprasegmental features of speech when assessing pronunciation proficiency. However, given the variation in approach and findings, it is difficult to make an assessment as to how such features should appear on rating scales. Constructing rating scales that incorporate suprasegmental features of speech is challenging because researcher measurements of what constitutes stress, intonation, and rhythm, whether instrumental or auditory, may not correspond to what raters and general listeners regard as constituting these suprasegmental labels. For instance, the study by Kang et al. (Kang et al., 2010) indicates that high rising pitch is influential to oral proficiency, but it does not follow that this term is suitable for a rating scale in the sense that raters will be able to interpret it in a meaningful and consistent manner. The following section examines pronunciation rating scales.

2.4 Frameworks and Rating Scales

This section of the literature review is concerned with summarising how several existing frameworks present aspects of pronunciation, introducing approaches to developing and validating rating scales, and reviewing references to pronunciation, and specifically, suprasegmental features, on existing scales.

First it is useful to define terms. A framework is the application of an abstract theoretical model of language performance to the assessment context (Fulcher & Davidson, 2007:36). Specifically, a framework is a document that relates the skills and abilities that a speaker is required to exhibit in a given context to achieve a given level of proficiency. It is not necessarily designed to be used for direct language assessment, but rather it is typically employed in the development of assessments and course materials. A rating scale, by contrast, is designed to guide raters in

assigning proficiency levels to speakers. It may be derived from a framework but developers are typically required to pay greater attention to issues such as usability in the assessment context.

2.4.1 Current Frameworks

American Council on the Teaching of Foreign Languages

The American Council on the Teaching of Foreign Languages produced a proficiency guideline in 1983, revised in 2012 (ACTFL, 2012). It is a 5-level set of holistic descriptors which focuses on what test takers can do in authentic interactions. Levis (2006) criticised the treatment of pronunciation in the previous iteration of this framework. He described it as “a haphazard collection of descriptors” and as “strikingly random in describing how pronunciation contributes to speaking proficiency” (p245 cited in Isaacs & Trofimovich, 2012). The problem persists in the current iteration of the framework. Parts of the ACTFL framework are devoted to pronunciation but references to specific features are vague and imprecise, e.g. the Advanced High level of the ACTFL framework refers to the use of “precise vocabulary and intonation to express meaning” (p5). Any further reference to suprasegmental features is only implied through terms such as “phonetic devices as a discourse strategy” (p5), which seems to be a reference to intonation, and the “oral paragraph”, which presumably equates to the paratone. Coverage of suprasegmental features in this framework is therefore limited. This is probably a reflection of the holistic nature of the framework. The guidelines are quite discursive and tend towards describing the types of interactions speakers would be successful in rather than the characteristics of successful delivery. For instance, Advanced Low speakers are regarded as:

“able to handle a variety of communicative tasks. They are able to participate in most informal and some formal conversations on topics related to school, home, and leisure

activities. They can also speak about some topics related to employment, current events, and matters of public and community interest.”

(ACTFL, 2012: 6).

Nonetheless, such limited detail on what constitutes pronunciation proficiency leaves a great deal open to interpretation.

Association of Language Testers in Europe

The Association of Language Testers in Europe (ALTE) produced a series of ‘can-do’ statements (ALTE, 2002) which provides functional descriptors for a range of different contexts. Similar to the ACTFL framework, these statements focus on the tasks a speaker can complete rather than the specific pronunciation features they may be able to deliver. Part of the reason for this is that the framework is supposed to be applicable across languages (Jones, 2000), and therefore it is not possible to describe how control of specific pronunciation features will be expressed. This framework makes no reference to suprasegmental features of speech.

Common European Framework of Reference for languages

Alongside the ALTE Framework a project was undertaken to develop the Common European Framework of Reference for languages (CEFR) (Council of Europe, 2001). This framework became “the standard reference document for teaching and testing languages in Europe” (Fulcher, 2004: 255) and has received much attention from researchers and test developers. It was developed by the Council of Europe in 1996, revised in 2000, and includes ‘can-do’ statements in a similar way to the ALTE framework and a ‘phonological control grid’ (Council of Europe, 2001: 117) devoted to pronunciation, which is reproduced in Figure 2.1.

Figure 2.1: The 2001 CEFR Phonological Control Grid

	PHONOLOGICAL CONTROL
C2	As C1
C1	<i>Can vary intonation and place sentence stress correctly in order to express finer shades of meaning.</i>
B2	<i>Has acquired a clear, natural, pronunciation and intonation.</i>
B1	<i>Pronunciation is clearly intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur.</i>
A2	<i>Pronunciation is generally clear enough to be understood despite a noticeable foreign accent, but conversational partners will need to ask for repetition from time to time.</i>
A1	<i>Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by native speakers used to dealing with speakers of his/her language group.</i>

(Council of Europe, 2001: 117)

The phonological control grid refers to suprasegmental features of speech with statements such as “can vary intonation and place sentence stress correctly in order to express finer shades of meaning.” (p117). However, attention to suprasegmental features is somewhat limited in this iteration of the CEFR, and broadly it has been met with some criticism. It has been described as:

“fully unrealistic when it comes to issues such as accent, or progression (particularly in moving from B1 to B2)”

and:

“not consistent as it mixes such diverse factors as stress/intonation, pronunciation, accent and intelligibility without providing clear indication of progression in any of these factors specifically”

as well as:

“not complete which results in jeopardizing its applicability and usefulness”

(Piccardo, 2016: 9)

Horner (2014) argued that the 2001 version of the phonological control grid relied too heavily on the concepts of ‘accent’ and ‘effort’. He proposes an alternative grid that progresses from “sufficient command of word stress to be understandable” at level B1, to sentence stress and intonation “are used successfully most of the time” (p119), as illustrated in Figure 2.2. It appears to present an intuitively sensible progression in control over suprasegmental features of speech, although, as with all frameworks reviewed here, effort is required to establish the finer details of performances at each level.

Figure 2.2: Horner's Proposed Phonological Control Grid

Table 2: Proposed new phonological control grid descriptors	
C2	<ul style="list-style-type: none"> • Speaker is easily understood. • Mispronunciations are rare. • Sentence stress is used successfully most of the time. • Intonation is used successfully most of the time.
C1	<ul style="list-style-type: none"> • Speaker is easily understood. • Mispronunciations are rare. • Sentence stress is used successfully most of the time. • Intonation is used but not always effectively.
B2	<ul style="list-style-type: none"> • Speaker is understood. • Mispronunciations occur but do not interfere with understanding. • Sentence stress is used but not always successfully. • Basic intonation patterns are used, but not always successfully.
B1	<ul style="list-style-type: none"> • Sufficient control of sounds to be understandable. • Sufficient control of word stress to be understandable • Mispronunciations occur, but only occasionally interfere with understanding.
A2	<ul style="list-style-type: none"> • Sufficient command of sounds and word stress to be understandable, but with some difficulty. • The interlocutor may need to ask for repetition or clarification.
A1	<ul style="list-style-type: none"> • Sufficient command of sounds to be understandable, but not all of the time and with some difficulty. • Sufficient command of word stress to be understandable, but not all of the time and with some difficulty. • The interlocutor will need to ask for repetition or clarification.

(Horner, 2010: 56)

A companion volume to the CEFR with new descriptors was released in 2017 (Council of Europe, 2017). This includes an updated phonological control grid which outlines the role of suprasegmental features in reference to three criteria: overall phonological control, sound articulation, and prosodic features. Stress, rhythm and intonation are elaborated in much more detail at all levels of this iteration of the CEFR. The descriptors, insofar as they refer to suprasegmental features, are reproduced in Appendix A. These descriptors present high proficiency speakers as being able to control stress, rhythm, and intonation to a fine degree, with speakers lower down the scale exhibiting less control and greater influence of their L1. This is a major improvement on the 2001 version of the scale and it covers suprasegmental features with a fairly high degree of detail.

Global Scale of English

The Global Scale of English (GSE) is framework that consists of descriptors each assigned a place on an 80-point scale. It contains details as to the suprasegmental features expected to be present. For instance, at the lower levels speakers “can use basic stress and intonation to support meaning” (Pearson, 2016a: 8), while higher on the scale “contributions are intelligible using intonation and stress to convey meaning, though this is not always consistent” (p9), and then a speaker “uses stress and intonation to convey subtle or inferential meanings”. The GSE is accompanied by Learning Objective documents which further populate the scale. These are ‘can-do’ statements describing performance at each point, for example, at point 65 on the scale speakers “can use intonation to indicate various degrees of certainty during a discussion” (Pearson, 2016b: 27). Among available frameworks the GSE and the 2017 update to the CEFR represent the best guidance as to the role of suprasegmental features in a learner’s progression.

2.4.2 Rating Scale Development and Validation

Rating scale descriptors define performances at different proficiency levels. In so doing they assign meaning to the resulting scores. Interpretation of test score meaning has been described as the primary concern for demonstrating test validity (Messick, 1996). Therefore, the construction of the rating scale is fundamental for the validity argument on which speaking assessments rest.

There are broadly three approaches to developing rating scales. Empirical approaches rely on instrumental analysis of speakers' speech signals, as well as statistical association between such features and rater judgement of proficiency (e.g. Brown et al., 2005; Fulcher, 1996; Fulcher, Davidson, & Kemp, 2011; Upshur & Turner, 1995). This establishes objective measures at each proficiency level as defined by raters. This procedure is advantageous because it references objectively measured elements of test taker performances and, as such, it reflects phonetic reality. However, it does not necessarily follow that raters will perceive such features as influential. Therefore, some developers opt to approach the construction of scales with reference to rater perception of proficiency (e.g. Brown et al., 2005; Ducasse & Brown, 2009; May, 2011; Orr, 2002; Pollitt & Murray, 1996). This has the advantage of relying specifically on the way raters attend to the speech signal. In contrast to instrumental approaches, however, it provides limited insight into what can be regarded as objectively present in the speech signal. The final broad approach to developing scales is measurement-driven approaches which can be used to scale descriptors regardless of how they are derived (e.g. North & Schneider, 1998). Rasch item response theory is employed to analyse the way in which raters use the descriptors, or to analyse rater judgements of descriptor level, to produce estimates of difficulty which can then be used to scale the descriptors. This has the advantage of being psychometrically robust and taking rater interpretations into account.

Scale Format

In its simplest form, a rating scale is a matrix which has criteria to be assessed as columns and the levels of proficiency as rows. The descriptors which populate the scale can be in a variety of formats and they are broadly described as being on spectrums between holistic and analytical, and impressionistic and objective (Harding, 2013). Pronunciation scales are typically impressionistic in that they require human interpretation. However, as reviewed above, there are speaking tests which are assessed using computer interpretations of the speech signal. More holistic scales broadly cover the performance without reference to specific fine-grained details and “express an overall impression of an examinee’s ability in one score” (Luoma, 2004: 60-1), whereas more analytical scales define in more detail the features of the performance required for it to be classified at a given level. The holistic approach has the advantage of being quicker and cheaper to implement and use because it may not require the same level of training and expertise to employ as analytical scales (Taylor & Galaczi, 2011).

The fine-grained approach taken by analytical scales is more common in high-stakes testing; this is due to there being less ambiguity in the scale points, making the rater’s process of reaching a decision more transparent and resulting in more reliable judgements. This seems to be the outcome in writing assessment (e.g. Weigle, 2002; Barkaoui, 2010), but it is less clear in speaking assessment which may be due, as Zhang and Elder (2014: 307) argue, to a range of factors. Speaking raters may be unable to align their intuitive judgements with the scale criteria, or they may struggle to sustain the higher degree of concentration required to attend to an analytical scale in the time-pressured environment of a speaking test. This may be more or less problematic depending on the number of criteria the raters have to assess and the context of the assessment. Analytical scales often rely on raters attending to more than one criterion simultaneously which may confound the measurement of each independent criterion. Equally, assessing two or three speakers in a discussion may be more arduous than measuring one in a monologue. In addition,

raters may lack the technical expertise required to analyse speaking against an analytical scale in contrast to a global one (Yates et al., 2011).

These issues, in combination with the 'underdeveloped' pronunciation construct (Galaczi et al., 2016: 157), contribute to a state where analytical scales are “highly problematic to design and implement” (Harding, 2016: 12). According to Kuiken and Vedder:

“establishing the relative weight of individual linguistic features in determining overall judgments of proficiency may often be problematic”

(Kuiken & Vedder, 2014: 331).

This is compounded by findings suggesting that raters are less confident in assessing pronunciation than other skills (Brown & Taylor, 2006; Isaacs et al., 2015), and have slightly lower confidence in judging features such as rhythm, stress, and intonation than global criteria (Yates et al., 2011: 14). A primary concern then for language testers, given the preference for analytical scales, is the extent to which human raters can attend to specific features of the speech signal in analytical detail. As Section 2.3 showed, it is not quite clear the degree of detail that raters can attend to stress, rhythm, and intonation and still be adequately reliable in their judgments. A resurgence in pronunciation research, as illustrated in Isaacs and Harding's (2017) review, demonstrates that defining and operationalising the pronunciation construct is an ongoing project.

Scale Validation

Validity means demonstrating that a test measures what it intends to measure. Early approaches to scale validation regarded validity as being separable into the follow types:

Content Validity: the extent to which the test samples the target-use domain.

Criterion-Related Validity: the extent to which test scores are related to other measures of proficiency or performance, either at the time the test is taken (concurrent validity) or in the future (predictive validity).

Construct Validity: the extent to which the test successfully measures the trait it intends to measure.

Face Validity: the extent to which the exam appears to be effective to stakeholders.

Response Validity: the extent to which the test takers and judges behave in a way expected by the scale developers.

(Alderson, Clapham, & Wall, 1995)

Two types of validity that are particularly relevant to this study are cognitive validity (Field, 2011) and scoring validity (Taylor & Galaczi, 2011). Cognitive validity is the degree to which the cognitive processes that are elicited by a test task mimic those that are exhibited in the target language use context. This is usually characterised as whether the task elicits the cognitive processes in the *test taker* that they will need when carrying out typical communicative tasks; however, it is also possible to consider cognitive validity in reference to the rater. Considering that intelligibility is constructed in part by the listener as well as by the speaker, it is not unreasonable to establish the degree to which a speaking test task elicits cognitive processes in the listener that are consistent with the process that will take place in typical listening outside the test centre. This is relevant to the discussion below in Section 2.5 on the characteristics that separate raters from typical listeners. Scoring validity is a broad term for the reliability of an exam. It encompasses the degree to which scores are “stable over time, consistent in terms of the content sampling and free from bias” (Weir, 2005: 23).

Messick (1989) applied a broad definition of test validity, arguing that it is the extent to which the interpretations of test scores are regarded as appropriate. This in turn was operationalised by concepts such as 'test usefulness' (Bachman & Palmer, 1996), and Kane's 'interpretive argument' approach (Kane, 2006), which seeks to track the inferences and assumptions made at each phase of administering and scoring a test.

Harding (2017) notes that currently there are limited validation studies relating to pronunciation because it is often rolled into broader validation of speaking, and also due to the under-representation of pronunciation in language-testing research generally. Nonetheless, he identifies several methodological approaches to validating pronunciation scales, including: correlations of features of speech with test score (e.g. Kang et al., 2010) or comparison of test scores with other measures of pronunciation proficiency (e.g. Bernstein, Van Moere, & Cheng, 2010); measurement approaches using Rasch item response theory (e.g. Isaacs & Thomson, 2013); experimental designs where individual elements of a test taker's performance are isolated (e.g. Crowther, Trofimovich, Isaacs, & Saito, 2015); focus groups (e.g. Harding, 2016); and written and verbal reports (e.g. Kim, 2009; Yates et al., 2011).

Reliability is fundamentally linked to overall test validity. It is conceptually inseparable from validity in the sense that if a test outcome cannot be replicated over different administrations of the test, all else being equal, then it stands to reason that the exam cannot be effectively measuring an underlying trait. Reliability then is a requirement of ensuring that the test measures what it is intended to measure. It bestows another important characteristic on tests in addition, one of fairness. Kunnan regards test fairness to be linked to validity in the following terms:

"The focus of this concern is on whether test-score interpretations have equal construct validity (and reliability) for different test-taker groups as defined by salient

test-taker characteristics, such as gender, race/ethnicity, field of specialization and native language and culture”

(Kunnan, 2000: 3)

In the domain of pronunciation assessment, native language represents a potential source of bias due to the complex interaction between speaker L1 and rater accent familiarity, as well as the varying influences of L1 interference on pronunciation. Rating scales that are not precise and understandable by raters are regarded as a risk to reproducibility in the case of writing assessment (Shaw & Weir, 2007: 172). A rating scale which does not precisely identify the features of speech that are to be assessed has the potential to unfairly punish test takers from specific language backgrounds, since it leaves raters to interpret the performances without adequate direction. The role of suprasegmental features is especially notable here since ‘correct’ articulation of some of these features, such as pitch choice, varies by characteristics such as dialect and speaker status (Gussenhoven, 2004: 72-3). The next two sections examine the amount of suprasegmental detail frameworks and scales exhibit.

2.4.3 Current Rating Scales

Rating scales developed by test providers also vary in the level of detail they contain. The International English Language Testing System (IELTS) exam includes an oral interview where candidates are assigned to one of nine bands on a range of criteria, including pronunciation. The descriptors for the pronunciation criterion of the public version of the IELTS speaking scale, reproduced in Appendix B, do not explicitly direct raters to suprasegmental features and include no reference to these, relying instead on the term “range of features” (IELTS, 2012: 19). The operational scale is rather more detailed and Galaczi et al. (2016) provide a useful hint by stating that the IELTS scale “includes descriptors such as ‘can sustain appropriate rhythm’ and ‘rhythm

may be affected by lack of stress-timing” (p178). It is probable then that the scale does include details of stress, and intonation.

Cambridge English has devised an overall speaking scale, reproduced in Appendix B, from which the operational rating scales used in the main suite exams (PET, FCE, CAE) are derived. The scale provides some detail on the place of intonation and word stress. For example:

CEFR B2:	Intonation is generally appropriate.
	Sentence and word stress is generally accurately placed.
CEFR C1/C2:	Intonation is appropriate.
	Sentence and word stress is accurately placed.

(Cambridge English, 2016b: 83)

This provides good information to raters on how much control over these features a speaker is likely to have. However, even with rigorous standardisation procedures, raters may struggle to separate ‘appropriate’ from ‘generally appropriate’ intonation. The scale also does not reference rhythm, although this is related to ‘sentence stress’, and is also likely to be encompassed in the Discourse Management criterion of the scale, where it requires test takers to produce “extended stretches of language with very little hesitation” (Cambridge English, 2016b: 83).

IELTS and the Cambridge English exams both use face-to-face oral tasks to assess pronunciation. The leading computer-based English speaking tests are the Test of English as a Foreign Language internet-based test (TOEFL iBT) and the Pearson Test of English (PTE). TOEFL iBT has two speaking scales for independent and integrated tasks, reproduced in Appendix B. These each include 5 levels, Level 0 (lowest) to Level 4 (highest). The highest level of the pronunciation (delivery) criterion on the independent scale states that speech “may include minor lapses, or minor difficulties with pronunciation or intonation patterns” (ETS, 2014). The level

below refers to minor intonation difficulties, which “may require listener effort at times (though overall intelligibility is not significantly affected)” (ETS, 2014). At Level 2 intonation is described as “awkward” and contains a reference to rhythm described as “choppy”. This scale therefore also appears to attend to intonation and rhythm to some degree of detail.

The Pearson Test of English (PTE) includes two tests for adults: academic and general English. The PTE Academic speaking assessment is made up of two 6-point scales which cover pronunciation and oral fluency. These are reproduced in Appendix B. The pronunciation scale includes reference to suprasegmental features at all levels, at the highest level “stress is placed correctly in all words and sentence-level stress is fully appropriate”, and at the lowest level “stressed and unstressed syllables are realized in a non-English manner” (Pearson, 2017a: 24). The oral fluency scale refers to rhythm at the highest two levels then to unevenness, hesitation, and staccato delivery and lower down the scale refers (Pearson, 2017a: 24). The operational test is machine scored and, although human raters provide raw data in the form of exemplar application of the scale from which the program derives scores, it is unclear exactly what role the scales play in that mode of scoring.

In summary, the status of suprasegmental features of speech on rating scales and frameworks is mixed. They tend to be missing from older frameworks of language proficiency, although the GSE and the new iteration of the CEFR refers to such features in detail. However, even where well-elucidated, there is still a great deal left for users to interpret. For example, the 2017 CEFR references L1 interference as damaging to control over intonation, but not all L1 interference is equally damaging, and foreign accent does not necessarily equate to less intelligibility (Munro & Derwing, 1995b).

Suprasegmental features of speech are treated with some precision on scales used by major exams. However, they are still rather broad, often relying on rater interpretation of what is

appropriate. Possible reasons for this are that developers are not able to ascertain where to put them. This is credible given the variation in research findings as to the influence of suprasegmental features (as described in Section 2.3 above), and the extent to which raters can reliably attend to them. Alternatively, developers may not believe raters are able to judge them effectively. Raters are the focus of the next section.

2.5 Raters

A rater's task is to judge a performance according to the set standard defined by the rating scale. Raters, therefore, can influence scores in "persuasive and often subtle ways" (Eckes, 2005: 198). The effect raters can have on scores is summarised by Taylor and Galaczi (2011: 209, in reference to Myford & Wolfe, 2003; 2004; and McNamara, 1996) as excessive variation in severity and leniency of scoring; a halo effect where scoring of one criterion influences scoring of an unrelated one; central tendency where raters neglect the extreme points on the scale; random variation; and bias where a characteristic of the test encounter influences scoring in certain circumstances but not in others.

What characteristics of the rater influence these rater effects? In the domains of speaking and writing assessments, researchers have identified several characteristics of raters that can influence their judgements. These include training (Davis, 2016; Barnwell, 1989; Hill, 1996; Shohamy, Gordon, & Kraemer, 1992; Weigle, 1998); English teaching and rating experience (Barkaoui, 2010; Lim, 2011; Isaacs & Thomson, 2013; Saito et al., 2016; Cumming, Kantor, & Powers, 2002; Elder, 1993; Weigle, Boldt, & Valsecchi, 2003), rater geographic location (Chalhoub-Deville, 1995; Chalhoub-Deville & Wigglesworth, 2005); and accent familiarity (Ballard & Winke, 2016; Browne, 2016; Carey, Mannell, & Dunn, 2011; Winke et al., 2011; Winke, Gass, & Myford, 2013; Winke & Gass, 2013).

These issues can be grouped under two broad categories: rating experience, which incorporates training and experience of English teaching and rating; and linguistic background. The latter also incorporates English teaching and rating experience due to such experience being obtained through interaction with non-native English speakers, as well as geographic location, which may cause variation in exposure to different accents in different degrees. Clearly these two categories are linked since, depending on location, experience of rating leads to exposure to different L1 speakers and hence promotes accent familiarity. The rest of this section reviews the influence of rater experience and accent familiarity on judgements.

2.5.1 Influence of Rater Experience

Speaking examiners are trained specialists who often have a background in language teaching and learning and have often undergone rigorous training and standardisation. They also often have considerable experience of listening to non-native speakers and learning another language. Nonetheless, it may be the case that non-experienced raters can judge language in a similar way to experienced raters:

“assessing communicative effectiveness is not an esoteric skill requiring arduous specialist training and licencing; it is one of the normal components of linguistic and social adulthood”

(Nichols, 1988: 14, cited in Barnwell, 1989: 6)

Yet analytical scales require that raters attend to relatively fine details of the performance, and speaking-test raters are typically not the same as the listeners a test taker may encounter in their language use context. This represents a risk to the validity of scores derived from such assessments. The role of expertise in these judgements is uncertain:

“it is unclear what particular linguistic training and experience is required for reliable judgments of complex linguistic phenomena in L2 speech (e.g. vowel reduction or pitch movement) and whether naïve listeners can achieve this”

(Saito, Trofimovich, & Isaacs, 2017: 442).

The variation in the way experience is defined and the criteria against which test takers are measured results in a complex landscape. Rating experience can be defined as simply the amount of time a rater has been judging language, but an experienced rater is also one who has undergone rater training and who is likely to have an extensive linguistic background. In assessment of writing, Eckes (2008) has been able to categorise experienced raters into six distinct types, but no such work has been undertaken in speaking assessment, except insofar as to distinguish between raters who tend towards global decision and those who tend to make analytical judgements (Pollitt & Murray, 1996).

Experienced raters have been found to be more lenient and more reliable than non-experienced raters in accentedness scores (Thompson, 1991). The study by Kim (2015) is illuminating in demonstrating that experienced raters assign scores more stably than non-experienced or developing judges, and that one training session was sufficient to improve the rating ability of the non-experienced and developing judges. However, this study is notable in its purely qualitative approach.

Other studies are less clear. Using many-facet Rasch measurement analysis, Davis (2016) demonstrated that additional experience of rating following training continued to improve rater consistency with reference scores, although reliability among raters did not improve as a result of additional experience. This indicates that experience has some impact on scoring, even if that is not an improvement in rater consistency. Additionally, several researchers have found non-

experienced raters to already be adequately reliable (Bridgeman, Powers, Stone, & Mollaun, 2012; Derwing, Rossiter, Munro, & Thomson, 2004). Trofimovich and Isaacs (2012) found 60 novice raters (undergraduate students from non-linguistic disciplines) made reliable judgments on 9-point Likert items for accentedness (Cronbach $\alpha = 0.99$) and comprehensibility ($\alpha = 0.99$). Non-experienced raters (graduate students from non-linguistics fields with no rating background) were found to not perform statistically differently from experienced raters (experienced ESL teachers) by Isaacs and Thomson (2013). However, Saito et al. calculated Cronbach's alpha to demonstrate that five applied linguistics graduate students were more reliable in assigning scores than five graduate or undergraduate students who were studying non-linguistic majors ($\alpha = 0.91$ versus $\alpha = 0.81$).

A complicating factor in this literature is variation in the way experience is characterised. Participants have been defined as experienced raters if they have experience of teaching English as a Foreign Language (Isaacs & Thomson, 2013; Trofimovich & Isaacs, 2012). In the case of Bongaerts et al. (1997), the experienced raters were teachers or phoneticians, and in Saito et al., (2016) they were teachers and graduates in applied linguistics. As a result, it is not clear what (if any) experience of examining in a formal oral assessment context the judges had. Notably, all the experienced raters in Trofimovich & Isaacs (2012) and two raters in Isaacs & Thomson (2013) reported undertaking no rater training; suggesting they did not have any formal experience of rating speech. Therefore, much of the research on experience actually focuses on language teaching experience or experience of studying language, rather than experience of making judgements.

On the basis of this evidence it may be the case that experience, however it is defined, results in an improvement in reliability of scores, even if non-experienced raters are already adequately reliable. This issue warrants further examination, though, because as it stands few studies define

experience as experience of professional language judgement, and none address pronunciation at the level of intelligibility.

2.5.2 Linguistic Background

Linguistic background has been described as a “proxy variable” (Han, 2016: 9) in speaking scale research because it takes the place of a range of other features associated with the rater’s experience of rating speech, or of working or living in a particular context or location. Testing organisations typically require speaking test raters to be experienced in teaching English as a foreign language before they are employed. The impact of this is that examiners are likely to have extensive linguistic experience, from communicating with non-native speakers in the classroom and often from living in foreign countries.

A rater’s exposure to the test taker’s L1 or L1 accented speech can introduce bias. Notably it can improve listener processing speed (Munro & Derwing, 1995b; Clarke & Garrett, 2004). For instance, Winke et al. (2013) investigated whether rater experience of learning a test taker’s L1 (Spanish, Chinese, or Korean) influenced the scores assigned on a 4-point holistic speaking scale. Using many-facet Rasch measurement analysis, they found that raters who had Spanish or Chinese as an L2 were more lenient when scoring speakers who had the same language as an L1. Equally, Carey et al. (2011) found that raters with extensive exposure to Cantonese, Korean, or an Indian language were more lenient to test takers with those respective L1s.

The picture is not completely clear though. Zhang and Elder (2011) found no difference in consistency or severity of scores by rating group when asking Chinese-speaking raters and native English-speaking raters to judge test takers on a 5-point holistic scale, although the qualitative element to the study demonstrated several differences in the way each group interpreted the

construct. Other studies have counterintuitive findings. Major et al. (2002) asked Chinese, Japanese, Spanish and native US-English speakers to take part in a listening comprehension test which included speakers from each of those respective L1s. The results were that some listeners comprehend speakers of their own L1 more easily (Spanish speakers), whereas others do not (Chinese speakers). The same irregularity with Chinese-accented speech manifests in objective measures of intelligibility. Bent and Bradlow (2003) used a key word transcription to find that Korean-accented speech was more intelligible to Chinese L1 speakers than either English-accented speech or Chinese-accented speech. It appears that speaker and listener background has a complex relationship to intelligibility.

Gass and Varonis showed that L2 exposure did not significantly affect a listener's ability to transcribe L2 accented speech (Gass & Varonis, 1984). However, when Kennedy and Trofimovich (2008) replicated the study with methodological issues resolved, they found that experienced listeners (EFL teachers) were significantly more accurate at a transcription task than listeners with minimal exposure to accented speech. They argue that the experienced-listener advantage stems from their experience of hearing a range of L2 varieties of English, which leads to them being better at decoding L2 speech. Browne and Fulcher (2016) found a similar improvement in intelligibility (measured via transcription) for raters who were familiar with Japanese listening to Japanese-accented English. This indicates that there is a difference between typical listeners, "the-man-in-the-street" in Tench's terms (1997, cited in Levis, 2006: 261), and experienced listeners in terms of intelligibility. Accent familiarity appears to influence intelligibility then, although there is limited evidence examining precisely how accent familiarity relates to intelligibility when measured on a rating scale.

It is intuitively sensible that familiarity with a speaker's accent would influence intelligibility, but researchers do not always link this to psycholinguistic theories of language acquisition, although

Browne is a relevant exception, deploying multiple trace theory to explain the influence of accent familiarity on intelligibility (Browne, 2016).

The articulation of words and utterances varies greatly from speaker to speaker, not just due to foreign accents, but even within the same speech community where people communicate with their own voices and their own idiolects. Yet even in the face of such variation, proficient listeners usually manage to adjust to new speakers of a known dialect easily.

Exemplar models of speech acquisition (Pisoni, 1997, Goldinger, 1997; Bybee, 2001: Chap 1; cited in Field, forthcoming) explain how listener familiarity with a particular sound system facilitates the decoding and interpretation of the speech signal of any speaker they encounter with the same accent. Multiple trace theory (Hitzman, 1986; Bybee, 2001) is one such model, proposing that exposure to various different speakers in different contexts leave traces of those encounters in the listener's mind, which can then be drawn upon in new encounters in order to find links between new and previous inputs. In other words, the theory posits that a listener does not hold one idealised version of a word to which they attempt to match the incoming signal. Rather they hold numerous versions of words said in different ways, in different voices, and in different contexts. Recognition then relies on the incoming signal being matched to these traces in memory.

The implications of this theory are that the details of a speaker's voice are encoded with phonetic information and stored in memory, rather than discarded once recognition has occurred (Pisoni, 1997: 11, cited in Field, forthcoming). In effect, this means that listener exposure to accents increases the ability to recognise speech delivered with such accents in future. A listener with more experience of hearing non-native speakers at various proficiency levels and in various contexts can be expected to retain more traces of such delivery in their long-term memory. In the

context of this study, an experienced teacher or speaking assessor is likely to find test-taker speech more recognisable due to retaining more traces of similar sounding words and utterances in similar contexts in their long-term memory.

2.6 Summary, Research Questions and Hypotheses

Three primary themes have emerged from the existing literature concerning suprasegmental features and intelligibility:

1. Intelligibility, in its narrow definition as perceptual recognition, is under-investigated insofar as it relates to rater judgements of proficiency. There is currently no research focusing on whether professional raters or typical listeners hold a common perception of the intelligibility of speakers who have limited control over suprasegmental features.
2. Certain suprasegmental features appear to influence pronunciation proficiency, both in studies correlating instrumentally measured suprasegmental features with pronunciation proficiency scores, and in studies collecting verbal report data. The findings are in no way conclusive, though, with methodological differences and varying researcher motivations resulting in inconsistent findings.
3. Rater experience appears to influence rater judgements generally and in reference to intelligibility specifically, when measured by transcription. However, the relationship between rating experience, intelligibility scores, and attention to suprasegmental features is under-examined.

The research questions that arise from this review of the literature are now stated and rationalised, after which there is a statement describing the original contribution to the literature that this thesis makes.

2.6.1 Research Questions

Research Question 1 (RQ1)

To what extent are experienced raters consistent in their judgements of the intelligibility of non-native speakers who exhibit suprasegmental errors?

Raters are less confident at assessing pronunciation, and particularly intonation, than other characteristics of a test taker's performance (Isaacs et al., 2015; Yates et al., 2011). Is this tentativeness warranted? The first research question addresses the extent to which there is agreement within a group of experienced raters as to the relative intelligibility of L2 speakers that exhibit different types of suprasegmental divergence from L1 norms. In so doing it identifies the degree to which experienced raters have a common perception of the intelligibility of such test takers.

By identifying the consistency with which raters interpret the intelligibility of suprasegmentally errant deliveries, this question establishes the initial suitability of such features to be examined by raters in an assessment context. Specifically, the inclusion of suprasegmental features of speech on rating scales depends on raters being able to judge them reliably. This is partly an issue of reproducibility, which is critical for resulting scores to have meaning, but also one of fairness. If the variation inherent in certain suprasegmental features causes irregular interpretation by raters, then the rater will represent a source of bias. Ultimately, these issues can be rolled up into the concept of scoring validity.

The question targets intelligibility for several reasons: 1) it is a foundation skill upon which comprehensibility is built (Nelson, 2008); 2) it relies primarily on perceptual clues and minimises the non-linguistic requirements of the listener (Smith, no date; cited in Nelson, 2008); 3) it

represents the ability of the speaker to code and the listener to decode phonological forms, which is grounded in psycholinguistic models of how the brain processes speech (Cutler & Clifton, 1999: 124); 4) it avoids reference to a native-speaker norm, which is difficult to define and does not reflect the reality of non-native interactions; and 5) it has been widely accepted in the existing literature.

Intelligibility defined as perceptual recognition has rarely been examined using a rating scale. Researchers typically use transcription tasks, for instance. By judging intelligibility on a 9-point scale, this question also examines such a scale for its efficacy in assessing pronunciation. Therefore, a necessary and important initial element of RQ1 is to examine the functioning of such a scale.

Hypothesis: Following research indicating that raters can reliably judge accent, comprehensibility, and fluency on a 9-point scale (e.g. Munro & Derwing, 1995a, Derwing & Munro, 1997), it was anticipated that experienced raters would be able to reliably judge intelligibility using the same format.

Research Question 2 (RQ2)

To what extent do the ratings of experienced raters differ from those assigned by non-experienced raters in terms of the scores they assign to suprasegmentally errant speakers?

Where RQ1 sought to identify whether experienced raters as a separate group have a common perception of the intelligibility of test takers, RQ2 is concerned with exploring the possible differences between these raters and typical listeners. When intelligibility is measured by a

transcription task, experience is known to promote intelligibility. However, it is unknown how far this extends to the language assessment context where judgements are typically made on scales rather than using objective tasks.

The differences between rater groups were examined in three ways:

- 1) the difference in rater reliability between groups, which establishes the degree to which each group has a common understanding of the intelligibility of the speakers. This makes it possible to infer a potential link between the interpretation of suprasegmental features of speech and rating experience. For instance, it may be the case that, due to experienced raters receiving training, they have a special propensity for establishing a common sense of intelligibility which is much more variable in the general population of listeners;
- 2) the difference in severity, which allows the link between rater experience and ability to recognise the speech signal to be examined. As discussed above, raters are a special case of expert listeners, and assessing the degree to which they differ from non-experienced raters in terms of severity establishes how much of impact this expertise has on judgements;
- 3) the difference in score range, which establishes whether the effect of being a trained, experienced rater has an influence on the range of severity that raters exhibit. Severity range among examiners is known to be resistant to training, so examining the range of severity further explores disparities between the ways experienced and non-experienced raters assess the intelligibility of suprasegmentally non-standard features.

The inclusion of a group of non-experienced raters addresses validity from two perspectives. It investigates the criterion-related validity of the intelligibility scale when used to measure

suprasegmentally non-standard deliveries. Specifically, it provides evidence showing the extent to which intelligibility to the raters in the language test can predict intelligibility to typical listeners. This speaks to the central purpose of many high-stakes speaking tests. For university entry, admissions policy often requires a minimum language requirement, not only so that students can read and write effectively in English, but also so that they can take part in seminars and group work, as well as deliver presentations to their peers. The other perspective from which this question targets the validity implications is to provide initial evidence as to the cognitive validity of rater judgements of intelligibility when listening to non-standard suprasegmental delivery. Finding that experienced listeners and non-experienced listeners regard test takers as equally intelligible would indicate that the way they interpret the intelligibility of speakers is similar, and as such, it would imply that they have similar cognitive processes in common. This is taken up in Research Question 3, where the actual characteristics of the signal that raters attend to is examined.

Hypothesis: On the basis of existing research, it was hypothesised that experienced raters would find speakers somewhat more intelligible than their non-experienced counterparts. Multiple Trace Theory, if it is accepted as the only currently convincing account of how listeners cope with highly variable speech signals, lends support to this assertion. In terms of reliability, non-experienced raters can be expected to be rather less reliable than experienced raters, due to their unfamiliarity with the procedure of rating and lack of prior training. Current research indicates that although non-experienced raters can be adequately reliable without any training, they tend to be less reliable than experienced raters. Finally, the range of scores was expected to be similar for both rater groups. There is limited existing research on which to base a hypothesis regarding whether one group would have a broader score range than another.

Research Question 3 (RQ3)

What suprasegmental characteristics of the speech signal do raters attend to when making judgements of suprasegmentally errant speakers?

This research question has two primary objectives: to ascertain the relative importance of different features as revealed in verbal reports; and to identify the emergence of what may be mutually comprehensible terms in the form of the actual words used by participants. The inclusion of both experienced and non-experienced raters allows for the degree of consistency between experienced raters and typical listeners to be examined.

Intelligibility is co-constructed by the speaker and the listener, therefore the rater's cognitive processes are an important element of speaking assessments. Identifying whether the suprasegmental features of speech that influence experienced raters complement those of non-experienced raters can provide insight into the cognitive validity of the assessment, in the sense that it may be possible to identify whether experienced raters think about a performance in the same way as 'typical listeners' a test taker may encounter in authentic interactions.

In studies addressing oral proficiency, it is common to state that there is a lack of empirical evidence indicating which features of pronunciation influence scores (e.g. Isaacs, 2008: 555), and gathering raters' perspectives on scales is rather less common than gathering the views of assessment experts (Harding, 2016: 14). This is certainly the case for suprasegmental features of speech. Yet in order to "convince others that the scores can be used to make inferences about learners to the knowledge, skills or abilities being tested" (Fulcher and Davidson 2007: 114), it is necessary to examine the precise interaction between a speaker's use of suprasegmental features of speech and rater interpretation.

There is strong support from instrumental studies for the influence of certain suprasegmental features of speech on pronunciation proficiency (e.g. lexical stress) but poor evidence for others (see Section 2.3). However, there is no guarantee that these findings can translate to the assessment context. For instance, it would not be possible to ask raters in an authentic speaking examination to attend to the specific influential tones identified by Kang et al. (2010) without giving raters extensive phonetic training, instrumental analysis tools, and multiple listens of the audio.

Perceptual studies provide promising insight, suggesting that raters may be able to attend to suprasegmental features in a great deal of detail, for instance the references to monotonicity and flat intonation found by Brown et al. (2005). However, studies which seek to identify how suprasegmental features are perceived by raters vary greatly in their findings. In certain studies, raters appear to not attend to suprasegmental features at all, even when prompted, whereas in others they discuss them in detail. Therefore, it is time for a study which examines rater perception of suprasegmental features when making intelligibility judgements.

Consistency in scores and feature attention opens the possibility to develop suprasegmental rating scale descriptors. Deriving descriptors from the language raters use when describing how they judge performances should result in descriptors that are usable. Additionally, usability may also be improved by comparing the language used by experienced raters to that of non-experienced raters, demonstrating whether rating experience endows listeners with a particular capacity to identify the features that influence their judgements.

Hypothesis: With the right analytical focus, listeners do attend to suprasegmental features of speech for certain criteria in certain contexts. Based on the discussion in Section 2.3, it was hypothesised that lexical stress and intonation would have an impact on intelligibility judgements. Lexical stress appears to influence intelligibility in laboratory studies using transcription, and in

judgements of ease of understanding. And intonation has been found to have a role in oral proficiency, broadly in instrumental and auditory studies, although intelligibility studies are mixed for intonation. Rhythm was also expected to influence intelligibility judgements since there appears to be a clear relationship between rhythm metrics and proficiency judgements, although notably there is very limited information on the role of rhythm in intelligibility. This may be related to the mixed findings of the role of focal stress in intelligibility. Misplacement of focal stress appears to not influence processing speed, and the findings of instrumental studies relating focal stress to intelligibility are mixed.

2.6.2 Original Contribution to the Literature

There are clear gaps in the literature in relation to rater interpretations of suprasegmental features. This thesis provides original contributions to current knowledge in several ways. By investigating raters' responses to speech, it provides empirical evidence for the place of suprasegmental features of speech on rating scales. It does this by establishing the extent to which raters can be consistent in their approach to judging speech that has suprasegmental deviations from norms. This is important because in human-scored assessments of speaking, raters do not typically use objective measures of the speech signal. In other words, rater perception is the primary way language is judged. Although several studies have examined raters' responses to the speech signal, none thus far has examined suprasegmental features of speech exclusively. As a result, this study can provide a much more exhaustive and thorough examination of the role of suprasegmental features of speech in the testing encounter. The outcome is an improvement in our current understanding of the pronunciation construct, as it is applied by language assessors. This is the first study to assess rater reliability when scoring intelligibility in its narrow definition as word and utterance recognition. Previously, intelligibility was typically measured using objective tasks such as transcription.

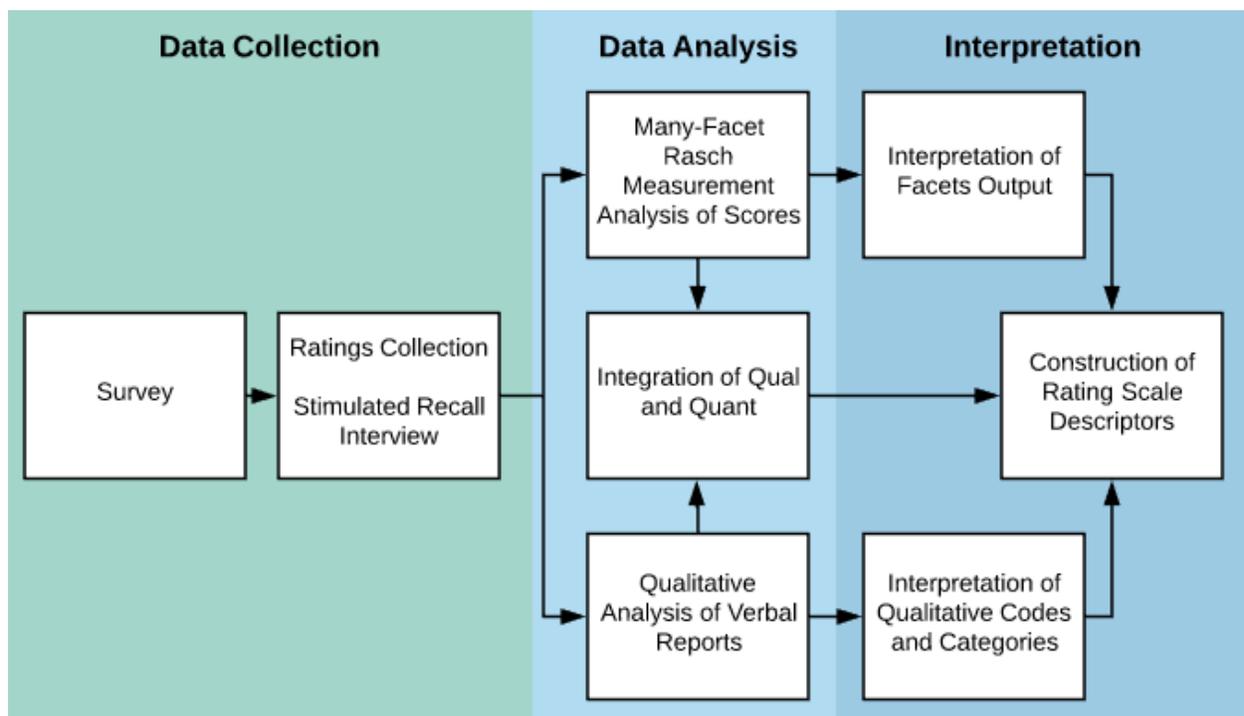
By examining the differences between experienced and non-experienced raters, this research provides information on criterion-related and cognitive validity. It examines the link between two very different populations, those who judge language professionally and have extensive experience with foreign-accented speakers and those who do not. Specifically, it establishes whether each population can reliably measure the intelligibility of speakers who exhibit suprasegmental deviations from norms. This has implications for the training and standardisation of raters. Raters typically do not receive specific phonological training, and this research will establish whether such training or standardisation is a requirement for avoiding inconsistency caused by suprasegmental features of speech.

The ultimate purpose of this research is to produce a series of rating-scale descriptors that reflect the way experienced raters interact with suprasegmental features. Using raters' own language in the development of such descriptors should promote their usability and overall face validity. Although a specific suprasegmental rating scale will shortly be released for the assessment of French speakers of English (Frost & O'Donnell, in press), no such descriptors have been devised for general English language assessment. This thesis, therefore, explores how suprasegmental features should be examined on pronunciation rating scales. Specifically, it explores the degree to which they must be represented analytically and impressionistically, as opposed to requiring objective computer-aided measurement.

Chapter 3: Research Design

A mixed methods design within a pragmatic theoretical framework was used to answer the research questions posed in the previous chapter. An approach mixing qualitative and quantitative methods was adopted because it allows the complexity inherent in judging speech to be explored more fully than in a single-method design. Specifically, both the process and the outcome of rating can be investigated using a combination of a stimulated recall procedure, and many-facets Rasch measurement analysis of the scores. Figure 3.1 below illustrates the research design which will be described in detail in this chapter.

Figure 3.1: Overview of the Design



This is a convergent parallel mixed-methods design (Creswell, 2014: 570) in which equal priority is given to both types of data in collection, analysis, and interpretation (i.e. QUAL + QUAN). The method of data collection is outlined in Section 3.1 and this is followed by a discussion of the participants (Section 3.2) and the materials used (Section 3.3). The pilot study is referred to throughout this chapter and summarised in Section 3.4, prior to a description of the ethical considerations, which are discussed in Section 3.5.

3.1 Method

This study is based on a pragmatic research framework. As a worldview, pragmatism does not conform to one position in terms of the nature of knowledge and reality. Johnson and Onwuegbuzie describe the usefulness of pragmatism thus:

“it offers an immediate and useful middle position philosophically and methodologically; it offers a practical and outcome-oriented method of inquiry that is based on action and leads, iteratively, to further action and the elimination of doubt; and it offers a method for selecting methodological mixes that can help researchers better answer many of their research questions”

(Johnson & Onwuegbuzie, 2004: 17)

The practical implications of this are that the research focus is on finding the best workable answers to the complex questions raised at the end of Chapter 2, rather than being restricted by a particular philosophical approach.

These characteristics make pragmatism well-suited to exploring the complex systems present in applied linguistics (Larsen-Freeman & Cameron, 2008). It is particularly apt for investigating the

constructed response tasks typically associated with testing productive skills. In these task types, the element of human scoring introduces a wide range of approaches and perspectives which may not be accessible through quantitative methods alone. In other words, the psychometric-structuralist tradition, which considers a quantitative approach to inter-rater reliability of scores to be of primary importance to scale validity (Broad, 2003: 5), may not be adequate, on its own, for validating speaking assessments. Human judgements of pronunciation have an element of subjectivity. A pragmatic approach allows enough methodological freedom for the requirements of psychometric validity to be examined, while also recognising that raters may not approach the task of scoring from similar perspectives.

3.1.1 Methodological Approach

This study employed a mixed-methods design, integrating quantitative analysis of scores with qualitative verbal reports from raters. Such an approach is valuable in providing “complementary insights” in pronunciation research (Munro & Derwing, 2015b: 13). The language assessment encounter is a complex one where raters make detailed judgements which are then distilled into a simple score. Mixing qualitative and quantitative methods is required to provide insight into both the rater’s decision making process, which is discursive and qualitative, and the resulting outcome, which is simple and quantitative.

The strength of the qualitative approach is to provide rich data on the complex thought processes raters undertake when making their judgements of intelligibility. It can identify precisely how raters respond to speakers who make suprasegmental errors. The strength of the quantitative approach is to provide generalisability, and to examine the issue in a format consistent with the context, in the sense that the assessment context relies on collection of scores. Both approaches provide insight into reliability, which is a critical issue in assessing language. By analysing scores, it is

possible to identify how consistent raters are in the outcome of the rating procedure. And by analysing qualitative responses it is possible to see how consistent raters are in the process of making their decisions. This is particularly powerful with the inclusion of non-experienced raters, as it allows the process and outcome of rating by professional raters to be contrasted with those by typical listeners.

A common alternative approach in this domain is a quasi-experimental design. This involves instrumental measurement of the features present in the speech signal and use of statistical tools to associate these features with scores (e.g. Galaczi et al., 2016; Saito et al., 2016; Pinget, Bosker, Quené, & de Jong, 2014; Galaczi, Post, Li, & Graham, 2011; Kang et al., 2010). This approach demonstrates the link between objectively measured features of the speech signal and rater-assigned scores. This project is primarily concerned with understanding rater perceptions and their interaction with the test taker, making a quasi-experimental approach unsuited to the aims. For instance, the attitudes and expectations that Hayes-Harb & Hacking (2015) found to be an influential factor in accent judgements would not have been uncovered without recourse to qualitative data.

Mixed-methods approaches are widely used in research within a pragmatic paradigm (Morgan, 2014: 1051) and they are a common tool for assessing rater behaviour (Yan, 2014; Hayes-Harb & Hacking, 2015; Zhang & Elder, 2014; Isaacs & Thomson, 2013; Kang, 2013b; Isaacs & Trofimovich, 2012; Ang-Aw & Goh, 2011; Rossiter, 2009; Iwashita et al., 2008; Hsieh, 2011).

Summary Overview

This research rests on a pragmatic research framework in the sense that it is concerned with finding the best solutions to the research questions raised in Chapter 2. Within this framework a

mixed-methods approach was employed which sought to identify both the outcome of the rating encounter, by gathering rating scale responses, and the process of the rating encounter, by gathering retrospective verbal reports.

A pilot study was undertaken where 10 raters judged the performances of 10 test takers on four 7-point scales, targeting comprehensibility, fluency, intelligibility, and overall pronunciation. They then reported how they assigned their scores in a stimulated recall procedure. Half of the raters were experienced in judging language and half had no such experience. Following the pilot, several amendments were made to the research design, which are discussed in the rest of this chapter and listed in 3.4 below. The main study was then undertaken with 30 raters who judged the intelligibility, comprehensibility and fluency of 12 test takers on 9-point scales. After which they reported their approach in a stimulated recall procedure. Half of the raters were experienced raters and the other half had no such experience. The performances were taken from Cambridge Main Suite exams.

3.1.2 Ratings Scale

Scores were collected to answer RQ1 and RQ2 by assessing whether raters had a common perception of the intelligibility of the performances, and how that perception varied based on experience. The rating sheet consisted of three 9-point ordered category items. The three criteria were comprehensibility, fluency, and intelligibility. This thesis is concerned only with analysing the intelligibility scores, but the comprehensibility and fluency items were included to ensure raters clearly separated these constructs from intelligibility. 'Overall pronunciation' was used as a criterion in the pilot. This was poorly defined and raters found it difficult to separate from the other criteria. Removing it for the main data collection therefore reduced the burden on raters when making their decisions.

On the rating sheet, a space was left for notes beneath each criterion. This facilitated the stimulated recall procedure by acting as an additional stimuli. The rating sheet is described in detail below in Section 3.3.2 and there is an example sheet in Appendix C.

A rating scale was selected as the instrument for measuring intelligibility for several reasons. Firstly, collecting numerical judgements allows the results to be analysed quantitatively, meaning rater behaviour can be examined in terms of consistency and severity. Secondly, such a scale replicates the way speech is measured in testing encounters, which promotes ecological validity. Thirdly, pronunciation proficiency criteria are commonly measured using rating scales in speaking test research; therefore, using a scale for the measurement of intelligibility promotes the commensurability of this study with others in the field. Fourthly, it avoids some of the problems associated with other ways of measuring intelligibility, such as transcription, in that it allows gradients of intelligibility to be recognised and reflects the different levels of processing required to recognise different qualities of non-native speech. Finally, it allows the qualitative data to be analysed alongside the score data in such a way as to develop a series of rating scale descriptors as described below in Section 3.1.4.

Munro & Derwing (2015b) suggest investigating intelligibility by using comprehension questions, summaries, True/False verifications, and identification tasks (p.14). It has also commonly been examined objectively by transcription (Gass & Varonis, 1984; Derwing & Munro, 1997). However, listeners are capable of putting linguistic constructs on linear scales (Southwood & Flege, 1999), and other studies have used ordered category items to reflect judgements of intelligibility (Llurda, 2000; Lumley & McNamara, 1995), although in these studies raters are asked simply to rate “intelligibility” without a clear definition of meaning.

3.1.3 Retrospective Verbal Reports

Verbal reporting is a procedure in which a speaker reports what they are doing or thinking as they are doing it or shortly afterwards. Verbal reports allow the investigation of explicit influences, i.e. those influences that raters are aware of and consider to be important. This is relevant to the speaking test context, where judges are typically required to make decisions by attending to certain rating scale criteria, without reference to instrumental analysis of the speech signal. The theoretical underpinning of this method of data collection is that:

“an individual’s verbalisations may be seen to be an accurate record of the information that is (or has been) attended to as a particular task is (or has been) carried out”

(Green, 1998: 1-2)

In other words, speakers can report what they were thinking at a given time with accuracy and veracity. This does not imply that the reports are complete accounts of the speaker’s cognitive processes. The verbal reports are raw data from which the researcher infers the speaker’s cognitive processes (Zheng, 2009: 127).

Concurrent reporting, where participants undertake the task and describe their thinking, was ruled out for this study because, as Ducasse and Brown (2009) note, speaking assessment takes place ‘on-line’, meaning the burden of listening and scoring is high and the introduction of an additional task may influence the target processes. Retrospective reporting is most effective when done immediately after the event under investigation (Ericsson & Simon, 1985: 19), so raters in this study provided their verbal reports immediately after assigning scores to all the speakers.

Verbal reports were selected to answer RQ3 because the resulting data reflects “participants’ own categories of meaning” (Johnson & Onwuegbuzie, 2004: 20). In order to promote this element of the method a ‘non-mediated verbalisation’ approach was taken (Green, 1998: 6), i.e. the verbal

report interview was unstructured and the researcher's prompts were designed only to encourage the informant to keep talking. This limits acquiescence bias since the researcher does not plant particular concepts in the minds of speakers, and all elements of the report are generated by the speaker.

A stimulated recall procedure is a specific way of collecting verbal report data where an informant's memories are prompted by stimuli. The presence of stimuli material improves results by jogging raters' memories (Someren, Barnard, & Sandberg, 1994: 27; Gass & Mackey, 2000: 17). The stimuli used in this study were the audio recordings of the test taker, the scores the rater had assigned to them, and any notes the rater had made on the scoring sheet.

Stimulated recall procedures have been used successfully in a number of studies relating to the rating of oral proficiency, and have been undertaken in several ways, including: 1) where raters listen to all performances once to score them, then listen again one by one and respond to questions in a semi-structured format (Hayes-Harb & Hacking, 2015: 60); 2) where raters watch a video of a performance, and then watch it again and pause it at points where they want to say something (May, 2011, Ducasse & Brown, 2009); 3) where raters watch a video of themselves rating and comment on what they were thinking at various points during the rating procedure (Winke & Gass, 2013).

A commonly cited challenge when collecting qualitative data relates to its veracity, the question of whether informants can and will report the truth. Reports may be incomplete due to raters not being able to retrieve all the relevant information (Crisp, 2008: 2; Brown, 2000), or raters may seek to confirm their own hypotheses, to appear consistent even in the face of contrasting information, or to use hindsight (Hoffman, Shadbolt, Burton, & Klein, 1995: 148). These issues were mitigated by carefully briefing the raters before collecting the data, as described below in

Section 3.3.3, and recognising the important step of researcher interpretation of the data in order to make judgements on reliability of responses using comparison and triangulation, as discussed in the next section, which addresses the approach to data analysis.

3.1.4 Analysis

The quantitative and qualitative data were analysed and interpreted separately, after which a strategy was employed to integrate them. Bazeley (2010: 432) considers the level of integration of methods in many mixed studies to be under-developed, and the current project resolves to challenge this by integrating the methods to develop a series of rating scale descriptors.

The quantitative data consisted of scores for intelligibility and were analysed using many-facet Rasch measurement analysis (MFRM) to answer RQ1 and RQ2. The qualitative data consisted of written notes on rating sheets and transcriptions of verbal reports, which were used to answer RQ3. Techniques associated with Grounded Theory were used to develop categories, and matrix queries were employed to analyse the interaction between raters, rater groups, test takers, categories, and codes. Finally, both forms of data were combined in the form of an integrated display to develop rating scale descriptors.

Many-Facet Rasch Item (MFRM) Analysis

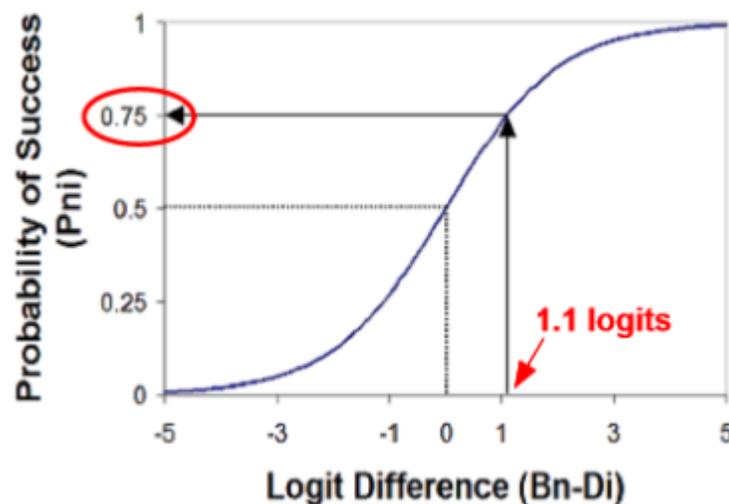
This section is concerned with rationalising the use of MFRM analysis. The specific details of the analysis undertaken in this study are reported in Chapter 4.

The Rasch model is prescriptive in allowing the behaviour of test takers, raters, and the scale to be analysed in reference to a pre-established one-dimensional latent trait, in this case 'intelligibility'. In other words, intelligibility is taken as a fixed trait which can be measured on a

single dimension. The extent to which rater scoring patterns fit the Rasch model reflects how closely the data corresponds to this assumption.

In the Rasch dichotomous model, where the data consists of zeros and ones (e.g. correct or incorrect answers), the model predicts the probability (i.e. the odds) of certain outcomes by analysing scoring patterns in relation to test takers and items. These odds are then converted into a common interval scale by expressing them as a logarithm. The units of the resulting log odds scale are known as 'logits'. The probability of getting the correct answer on any given item can be predicted on the basis of a test taker's logit score. A higher logit score reflects greater ability in the latent trait. Figure 3.2 below illustrates this relationship, showing that if there is no logit difference in the ability of the test taker and the difficulty of the item, the probability of the test taker getting the correct answer is 0.5. If the test taker's ability is 1.1 logits higher than the item difficulty, this increases the probability of getting the correct answer to 0.75.

Figure 3.2: Logit Difference Between Test Taker Ability and Item Difficulty Against Probability of Success



(Linacre, 2012)

In mathematical terms, the dichotomous model predicts the probability that person n , of ability B_n , succeeds on item i , of difficulty D_i .

$$\log_e \left(\frac{P_{ni}}{1 - P_{ni}} \right) = B_n - D_i$$

(Linacre, 2012)

This model has been expanded to take account of rating scale data (Andrich, 1978). In the rating scale model, the scale is conceptualised as a series of dichotomies between higher and lower categories. In mathematical terms, this rating scale model predicts the probability that a person n , of ability B_n , is observed in category j of a rating scale applied to item i , of difficulty D_i .

$$\log_e \left(\frac{P_{nij}}{P_{ni(j-1)}} \right) = B_n - D_i - F_j$$

The rating scale $\{F_j\}$ is the same for every item

(Linacre, 2012)

Where the dichotomous model is concerned with the probability that a candidate of a certain ability will correctly answer an item of a certain difficulty, the Rating Scale model is concerned with the probability that a candidate of a certain ability will be assigned to a rating scale category of a certain difficulty.

A further expansion of the Rasch model is many-facet Rasch measurement analysis, which allows additional variables beyond item, rater, and test taker to be included in the analysis. For instance, in addition to measuring the probability associated with a certain test taker receiving a certain score from a certain rater, the model can take account of how raters or test takers with a given characteristic are likely to behave.

This method of quantitative analysis was selected for use in this project because it allows individual rater performances to be observed, for example in terms of severity. By gathering scores and applying the “magnifying glass” of many-facet Rasch measurement (Sawaki, 2007: 357), it is possible to identify whether the scale and raters work effectively to provide a valid estimate of intelligibility. Raters whose perceptions of intelligibility deviate from the norm established by the rest of the rater group can be singled out for investigation in more detail, thus complementing the qualitative analysis. In other words, the quantitative score collection and analysis provides insight into the outcome of the rating in such a way that raters and test takers can be compared and contrasted on a common scale.

Qualitative Analysis

The purpose of this section is to rationalise the use of qualitative analytical techniques. Details of the transcription, coding and analysis of the qualitative data are reported in Chapter 5.

The qualitative analysis employed components of the Grounded Theory Method (GRM) (Glaser & Strauss, 1967). Urqhart et al. (2010) identify the primary characteristics of this method as: 1) a concern with developing theories; 2) limiting the preconceived ideas the researcher brings to the data; 3) the process of constant comparison; and 4) sampling driven by theoretical concerns. It is common for researchers to modify GRM to suit the needs of their particular research aims (LaRossa, 2005), and this is the case in existing research relating to rating scales (e.g. Wei & Llosa, 2015). One of the primary elements utilised in the current study is open coding, which is an inductive system of coding that does not rely on a pre-existing code book but rather allows codes to emerge from the text. The other element is constant comparison, a process whereby data is coded iteratively, with new codes and categories compared to previously defined codes

and categories throughout the coding and categorisation of the data. Both of these techniques are discussed in practical terms below in Section 5.1.2

The purpose of using these techniques was that they are in line with the aims of the research to derive categories and codes that reflect the way raters attend to the speech signal in their own terms. The initial round of coding was not restricted by preconceived ideas or theoretical assumptions as to the way in which raters would engage with the speakers. As a result, the codes represent the raters' own ways of describing the performances. Suprasegmental codes were not imposed on the data from the start, as the focus was not brought to suprasegmental features until all the data had been open coded.

The specific technique used to answer RQ3 was to systematically compare codes and categories using the matrix function in NVivo 10 (QSR International, 2012). This function allows codes and categories to be compared in terms of counts, and in terms of the relevant parts of the transcriptions. Raters were compared within each group and between rater groups, in order to understand how experienced and non-experienced raters attended to the speakers in suprasegmental terms and what the differences were between raters on the basis of experience. Using matrix queries allowed order to be brought to the data from a broad focus, such as when observing the relative attention all raters gave to specific categories, to a very narrow focus, such as observing a specific rater's approach to a specific speaker with reference to a specific feature.

Integrated Display

An integrated data display is a way of showing the relationship between qualitative and quantitative data by placing both types of data on a common visual display (Onwuegbuzie & Combs, 2010: 423). Plano Clark et al. (2010) describe it as 'merging with a matrix' because, in

practice, an integrated display is a matrix with quantitative categories as rows and qualitative categories as columns. This is the same structure as a typical rating scale, making it particularly suitable for placing candidate performances against scores, as demonstrated by Lee and Greene (2007).

An integrated display can be constructed following the analysis of qualitative and quantitative data. It requires another round of analysis relating quantitative groups with qualitative categories. The construction of the integrated display is discussed in detail in Chapter 6 but, in summary, the columns of the matrix were named after the suprasegmental categories that emerged from the qualitative analysis and the rows were named after the moderated scores from the facets output. The qualitative data was then grouped according to the quantitative categories and the integrated display was populated by observing the way suprasegmental categories were expressed in each group.

3.2 Participants

Two groups of participants were recruited for data collection: experienced raters and non-experienced raters. Participants were recruited using a nonprobability sampling strategy (Teddlie & Yu, 2007), where raters were purposefully targeted based on their characteristics, i.e. rating experience.

As noted above in Section 2.5, a variety of differing criteria have been used to define rater experience. For this study, experienced raters are defined as those who had received item writer training from an established examination board and who had examined speaking tests for at least one year. Experienced raters were recruited for this study by two routes: the Cambridge Assessment Speaking Team sent out a call for participants to registered UK-based examiners;

and the researcher used his own personal and professional contacts. Additional experienced raters were recruited using word of mouth.

Non-experienced raters have been defined as those with no training in language or linguistics (Bongaerts, Van Summeren, Planken, & Schils, 1997), with no linguistics or teaching experience (Callaway, 1977), with no experience of teaching an L2 or of undertaking specialised training (Isaacs & Trofimovich, 2012), and with no assessment background (Isaacs & Thomson, 2013). In the current study, non-experienced raters are defined as those who have never undertaken rater training and have never examined language tests. Additionally, to control for accent familiarity, the non-experienced raters were people who had never lived outside the United Kingdom, had no experience since secondary education of learning a foreign language, and did not have any close friends or relatives who spoke English as a second language. Non-experienced raters were recruited through the researcher's own contacts and word of mouth.

The sample size was influenced by two factors. In terms of the quantitative element, Linacre (1994) states that for usefully stable Rasch item response theory calibration, a sample of 16 – 36 is required. Qualitative data is recognised as often being more time consuming to collect and analyse than quantitative data (Johnson & Onwuegbuzie, 2004), and the resources required to collect and analyse the qualitative data represented the biggest restriction on sample size. A total of thirty participants satisfies the minimum sample requirements of the quantitative analysis while providing a large enough amount of qualitative data to reflect a range of views. This sample size is consistent with other similar mixed-methods studies that employ Rasch IRT (e.g. Kim (2009) had 24 raters; Yan (2014) had 11; Isaacs & Thomson (2013) had 40; and Zhang & Elder (2014) had 38).

All prospective raters took part in a questionnaire, administered via the Survey Monkey website (www.surveymonkey.com). The primary purpose of the survey was to efficiently identify which people were suitable to take part in the data collection. The survey took the form of a structured questionnaire of closed and open questions. Reliability was supported by ensuring the questions were straightforward and left no room for participant interpretation. A quick and convenient online survey supports the ethical responsibility of the project by limiting the amount of time required of potential participants. Prospective raters were deemed suitable if:

1. they identified themselves as native English speakers. No other test of nativeness was applied, although in the event all raters who took part had been born and spent their formative years living in countries where English was the first language. Nativeness was controlled because some studies have found that native speakers and non-native speakers score differently (e.g. Hill, 1996). Although notably other studies have found that they are consistent (Kim, 2009). In Zhang and Elder's (2014: 308) literature review on the subject, they conclude that the influence of rater L1 on scores is inconclusive. Therefore, it was deemed appropriate to limit the raters to native English speakers and to accept the resulting reduction in generalisability of the findings.
2. they were aged 18 - 60. People under 18 are unlikely to be employed as language raters, and aging is potentially related to decline in listening skills (Schneider, Daneman, & Pichora-Fuller, 2002).
3. they did not have a hearing impairment.
4. they could be clearly categorised as experienced or non-experienced according to the following criteria:

- | | |
|-------------------------------|--------------------------------------------------------------------------------|
| Experienced raters | i) had been trained by an established exam board, such as Cambridge Assessment |
| | ii) had at least one year of rating experience |
| Non-experienced Raters | i) had never taught or examined languages |

- ii) had not studied a language other than English since secondary education

Table 3.1 below illustrates the demographic breakdown of raters. It shows that there were more female participants than male and that the mean age of non-experienced raters was slightly higher than experienced raters. There was a broad range of examining experience within the experienced raters group, with the most experienced rater having worked as an examiner for 35 years and the least experienced having done so for two years.

Table 3.1: Summary of Raters

	Experienced	Non-Experienced	Total
Female	9	10	19
Male	6	5	11

Mean age	47.04	47.18	47.1
Minimum age	29	25	25
Maximum age	59	55	59

Mean examining experience (years)	9.38
Minimum experience (years)	2
Maximum experience (years)	35

Sixty-three per cent of raters were female. There is considerable research into how the difference in rater styles during the interview interaction influences test taker performances (e.g. see Reemann, Alas, & Liiv, 2013), but research into whether rater gender influences scoring in oral tests is inconclusive: “most of this research reveals some kind of gender effect on test scores although, interestingly, the effect is not always the same” (O’Loughlin, 2002: 171). There were far more female respondents to the call for research than male. This imbalance made recruiting raters

balanced by gender difficult, although it is consistent with other research requiring Cambridge Assessment speaking examiners. For instance, 89% of the respondents to Gilbert and Staub's (2014) survey of speaking raters were female. It is not clear whether this greater proportion of female to male raters reflects the balance of raters in typical language assessments.

The experienced raters all spoke one or more foreign language, ranging from one rater who reported elementary French to another who was proficient in Russian, German, Czech, and Italian. They also all reported extensive exposure to many varieties of non-native English speech due to working as English teachers in multilingual classrooms and examining test takers in examinations. Exposure to a range of foreign accents on the part of raters contributes to the study's ecological validity, i.e. familiarity with a speaker's L1 is 'construct relevant' (Browne & Fulcher, 2016: 51) because raters do come from a wide range of backgrounds with a wide range of linguistic experiences. The range of experience and L2 ability represented in the study is typical of a group of UK-based professional raters.

The examiners classified as experienced had experience in a range of exams. Table 3.2 below demonstrates that Cambridge Main Suite was the most commonly reported exam.

Table 3.2: Rating Experience

Exam	Number of raters with experience examining
Cambridge Main Suite	10
IELTS	7
Skills for Life	5
TOEFL	4

Note: several raters examined for more than one exam board

It is unsurprising that Cambridge Main Suite exams were the most common source of rater experience due to the majority of participants in this group being recruited via Cambridge Assessment.

3.3 Materials

Data collection was facilitated using audio recordings of authentic tests and a rating sheet. The development of these materials is now described and rationalised, and this is followed by a description of the data collection procedure.

3.3.1 Audio Samples

Raters were asked to judge test takers on the basis of audio-only stimuli rather than in a face-to-face encounter. This is common in pronunciation proficiency studies and does limit the influence of non-linguistic factors, such as facial expression, but it is not consistent with most speaking test encounters. Video was a possible alternative to audio only, but it still does not exactly replicate the face-to-face exam encounter and introduces a range of other issues, such as sourcing video stimuli that is consistently adequate quality. The limited research into the difference in judgements

between these two conditions suggests that raters may be better able to discriminate among candidates if video were available (Pan, 2015: 10). However, more research is needed to understand what impact, if any, the use of video or audio-only modes of assessment have on pronunciation rating.

Recordings of exams were provided by Cambridge English Language Assessment. This constituted 19 pairs or sets of three test takers completing PET, FCE, or CAE exams. In total, 42 speakers were included in the recordings. PET, FCE, and CAE exams were the focus because they are linked to levels B1 - C1 on the CEFR (Council of Europe, 2001). Exams at this level of competence were chosen because B2 represents the point at which speakers can “interact with a degree of fluency and spontaneity... without strain” (Council of Europe, 2001: 24), so it is likely that test takers will still be overcoming suprasegmental challenges as they approach and exceed this point (Field, 2011). Additionally, Galaczi et al. (2016) found that between CEFR levels B1 and B2, speakers gained better control over reduced vowels, indicating that this is a relevant area to target when examining rhythm and stress.

The long turn monologues were extracted from each recording for analysis. Monologues were chosen over dialogues for practical and methodological reasons. Practically, it is difficult for raters to distinguish between two speakers of the same gender in an audio recording, thus adding a level of the complexity to the task. More importantly though, characteristics of the test taker’s interlocutor, such as personality and degree of acquaintanceship, can influence the quality of speech (Beebe, 1980; Coates, 1993; O’Sullivan, 2002; Wolfson, 1989; cited in Galaczi & French, 2011: 166). Therefore, when examining specific features of speech, it is necessary to control factors which may provoke variation in the test taker’s control over these features. The choice of monologues does restrict the pronunciation features that can be exhibited, and this limitation is discussed in Chapter 7.

The researcher reviewed all 42 monologues and identified speakers who, based on the researcher's impressionistic judgement, deviated substantially from suprasegmental English L1 norms. A number of resources were used at this stage to define an L1 norm. *Learner English* (Swan & Smith, 2001) provided an excellent overview of the issues faced by speakers of various languages. Additionally, the errors typically addressed in language learning text books, such as the *English File* series (Oxenden & Latham-Koenig, 2013) provided resources to identify typical errors a language learner will make. The following features were examined at this stage:

1. Stress (pitch movement, duration, and relative loudness).
 - a. Lexical stress: Relative prominence of syllables within a word (Giegerich, 1992).
 - b. Focal stress: Relative prominence of syllables within an intonation unit (Ladd, 2008).
2. Rhythm: perception of a consistent pattern of relatively strong and weak events in the speech signal (Ashby & Maidment, 2005: 161).
3. Intonation: pitch movement over word groups (Laver, 1994)

The researcher sought to select audio samples that not only illustrated a representative range of suprasegmental issues, but were also controlled to retain a balance of gender, level of English proficiency, and L1. Following this initial review 14 clips that met the above criteria were selected. The initial analysis undertaken by the researcher is presented in Appendix E.

These clips were then submitted to a panel of four phoneticians for review. The members of the panel were instructed to independently describe the suprasegmental features that characterised these speakers. In doing so, they validated the researcher's choice of the recordings and provided a means for selecting specific speakers for use in the study. This ensured that suitable, well-

attested examples of a range of different suprasegmental errors were present in the speech samples. The panel of phoneticians consisted of:

Dr Yuni Kim, University of Manchester

Dr Richard Cauldwell, Independent researcher

Dr Linda Shockey, University of Reading

Dr Rachel Smith, University of Glasgow

The procedure for developing the final audio rationale summaries was to present each phonetician with the detailed rationales, in addition to the audio clips, and ask them to comment on the veracity of the rationale. When the comments were collated, the result was a short summary of the way in which each speaker's suprasegmental delivery was not standard. Sometimes the phoneticians agreed with the researcher's rationale, however, they often had suggestions for ways to be more precise, and to identify features that had been neglected. For instance, Speaker 5 was initially considered to have limited rhythm and Dr Cauldwell refined this by arguing "the big thing for me is that [she] is monotonic in her delivery". Equally, Speaker 4 was initially regarded as applying incorrect emphatic stress. Dr Shockey stated that this was not quite accurate and the perception probably came about due to the rising intonation at the end of many phrases, giving the impression of emphasis. The resulting summaries, presented in Appendix F, represent the aggregate agreement of the four expert listeners and the researcher as to the suprasegmentally notable features in each performance.

The 14 clips were then reduced to 12 to reduce the total time required for data collection so as to ensure that scores and verbal reports could be collected in less than 90 minutes. This avoided making unrealistic demands upon the attention span of the raters who were asked to assess the samples. The two speakers removed were those deemed to exhibit errors that were already well represented in the sample of test takers.

A primarily auditory approach, rather than an instrumental approach, to selecting suitable stimuli was adopted because the context of the study rests on perceptual recognition of the features present in the speech signal. Instrumental analysis is not able to assess a particular feature's impact on listeners. This is because instrumental phonetic abstractions of pitch (such as fundamental frequency) are not necessarily aligned with human experience of pitch (Cruttenden, 1997: 5-6). Furthermore, in practical terms, pitch can be hard to detect instrumentally, not least because vocal folds do not vibrate during voiceless consonants, but also because the technology requires well-enunciated speech, good quality recording equipment, and a quiet environment. Wennerstrom (2001) argues that instrumental approaches require unnaturally exaggerated speech. She states: "I tried in vain to obtain a computer pitch reading for the word *fact* uttered in fast speech" (p33), highlighting the difficulties of identifying intonation by computer.

This holistic approach to selecting stimuli material has the advantage of taking a broad view of suprasegmental features. The question over whether segmental or suprasegmental features of speech are more important to intelligibility is misleading since different elements of segmental and suprasegmental phonology may be important in different magnitudes (Celce-Murcia, Brinton, & Goodwin, 2010). Furthermore, the distinction between them may not be valid (Zielinski, 2015), for example, as mentioned above a potentially critical element of the perception of lexical stress is vowel reduction, which is segmental. Zielinski (2015) concludes that future research should focus on the "cumulative effect of multiple non-target-like features" (p407) rather than the relative importance of individual features.

Table 3.3 summarises the final 12 audio files. They ranged across three exams: Cambridge English: Preliminary (PET); Cambridge English: First (FCE); and Cambridge English: Advanced (CAE). They represented a broad range of L1 speakers. Importantly for the aims of this study there was a range of languages regarded as having rhythm that is stress-timed (e.g. German

[Kohler, 1982]), those regarded as having rhythm that is syllable timed (e.g. Spanish [Pike, 1946]) and mixed (e.g. Korean [Seong, 1995]). There is also a mix of tone languages (Mandarin, Vietnamese) and non-tone languages (e.g. Russian, French).

Table 3.3: Summary of Audio Stimuli

Number of clips: 12		
Mean length of clips: 56 seconds		
Exams	PET	5
	FCE	4
	CAE	3
Gender	Female	6
	Male	6

L1	French	1
	German	2
	Italian	1
	Korean	1
	Mandarin	2
	Russian	1
	Spanish	2
	Vietnamese	2

On the basis of existing research it seems to be the case that raters are able to make reliable judgements of spoken English with relatively short audio clips, as little as 10 seconds in Schmid and Hopp’s (2014) study, and approximately 18 seconds in Browne and Fulcher’s (2016). Nonetheless, some raters during the pilot study felt that the length of exposure to the performance was not sufficient to make their decisions. As a result the mean length of the audio clips used in the final study was 56 seconds (min: 47s; max: 69s). This is 33 seconds longer than the average audio clip used in the pilot study.

Another addition to the main study which came about due to concerns raised during the pilot was the inclusion of a standard setting procedure. Pilot raters reported being unsure where to place the first few test takers on the scale and wanting to make changes to scores after finishing the scoring due to being aware of the broad ability level of everyone involved. Without a standard setting procedure, it was difficult to know how to use the rating sheet, and raters unavoidably compared later test taker performances to the earlier ones. This is supported by research

suggesting “the basic psychological process of magnitude judgment is one of comparison” (Davis, 2016: 119). Therefore, to facilitate the standard setting, three additional audio clips were selected. These audios represented a low, medium, and high ability level and were chosen on the basis of the original pronunciation scores assigned by Cambridge Assessment examiners. Before beginning the rating procedure, raters were asked to listen to these three audio clips and were advised that they represented approximately the range of ability that raters would hear during the rating.

3.3.2 Rating Sheet

The rating sheet (see Appendix C) included three 9-point ordered category items designed to measure comprehensibility, fluency, and intelligibility. The purpose for including comprehensibility and fluency was to ensure raters clearly separate these criteria from each other. Below each item there is space for raters to write notes, which were then used as part of the stimuli during the collection of verbal reports.

During the pilot study, the audio stimuli were not presented to the raters in a contextualised format, and they did not know what exam prompt the test taker was responding to. Some raters reported feeling that they were spending time trying to establish the context of the speech rather than simply judging the performance. To address this, a context sentence which outlines the exam prompt was printed on each rating sheet for the main study. This improves context validity since authentic examiners would, of course, know the exam prompt and familiarity with the topic of discussion may improve intelligibility (Gass & Varonis, 1984).

In order to avoid raters applying their own interpretations, the terms ‘comprehensibility’, ‘fluency’, and ‘intelligibility’ were not used on the rating sheet. Instead they were transformed into questions

which represented the definitions used for this study. Failing to clearly describe the constructs under investigation to raters causes issues with rating (Jun & Li, 2010: 61), and, as discussed in the literature, there are a range of definitions for intelligibility. Table 3.4 below shows how each criterion was described on the rating scale and the source of that definition in the literature.

Table 3.4: Rating Scale Criteria

Criterion	Rating scale question	Source of definition
Comprehensibility	How easy is it to understand the speaker?	Derwing et al., 1998: 396
Fluency	How well does the speech flow?	McCarthy, 2010: 12
Intelligibility	How easy is it to recognise the words and phrases?	Smith & Nelson, 1985

For the intelligibility criterion, the somewhat technical term ‘utterances’ was changed to ‘phrases’ in order to improve clarity. In line with standard practice in pronunciation research, textual descriptors were provided at the extreme points of each item (Isaacs & Thomson, 2013: 141; Knoch, 2016: 55): ‘Very Easy’; ‘Very Difficult’; ‘Very Well’; ‘Very Badly’. These textual descriptors provide an anchor on which raters can position each speaker’s performance. Absolute end points (e.g. ‘extremely’) were avoided to reduce the risk of central tendency bias.

Nine-point ordered category items were used because this many points was deemed to provide sufficient sensitivity to reflect a range of rater perceptions in fine detail. Isaacs and Thomson (2013) considered 9-point items to be “practical, usable across contexts, and sufficiently reliable for research purposes” (p136). The pilot study used 7-point items and there was a tendency towards a ceiling effect, a known risk of scales shorter than 9-points (Southwood & Flege, 1999). Nine-point items are common in pronunciation proficiency studies (e.g. Saito, Trofimovich, &

Isaacs, 2016; Munro & Derwing, 1994; Munro & Derwing, 1995b; Munro & Derwing, 1999; Rossiter, 2009; Isaacs & Trofimovich, 2012; Jun & Li, 2010), which adds an element of commensurability to the findings of this study.

3.3.3 Procedure

Once raters had completed the survey and it was established that they fit the criteria to take part in the project, they were invited by email to meet for the data collection. The researcher met each participant individually. The quantitative and qualitative data were collected consecutively in a single sitting. Each session lasted between approximately 50 minutes and 80 minutes depending on how much time a rater spent making their verbal report.

Initially raters listened to three exemplar responses to set the standard. Following this, raters were asked to listen to each test taker and assign a score for the three criteria. Finally, they took part in a stimulated recall procedure where they described what they were thinking when giving the initial scores. What follows is a detailed description of the three stages of data collection.

Phase 1: Standard Setting

Firstly, raters were asked to read and sign the informed consent form. Then the researcher summarised the procedure highlighting the important points, such as the focus on pronunciation, and the centrality of recalling memories for the stimulated recall procedure rather than formulating new insights.

Raters were then presented with the scoring sheet and asked to familiarise themselves with it. The researcher verbally described and indicated each element of the sheet. They were asked to read each criterion and confirm they understood it.

Finally, raters were advised they would hear three thirty-second clips, each one defined as low, medium, or high ability. Raters were advised that the low candidate in the clips could be expected to be the similar to the weakest speaker they would hear, the medium candidate would be a good reflection of the middle ground in terms of the relative ability of the candidates, and the high-ability speaker to reflect the higher ability level of the test takers. They were advised that these categorisations might not reflect specific scores for comprehensibility, fluency, and intelligibility but were meant as an indication of the general pronunciation ability they could expect to hear throughout the rating exercise.

The audio clips were saved on a Dell laptop running Windows Media Player and played through Bose headphones.

Phase 2: Score collection

Raters were given one rating sheet for each audio clip. The sheets were ordered randomly to reduce ordering effects and raters were asked to work through them one by one. Besides the context sentence, raters were not provided with any other information which might influence their judgements, such as the age, gender, or L1 of the test takers. The researcher gave raters an opportunity to ask questions before proceeding.

The researcher played each audio once, giving the rater as much time as needed to assign a score for each criterion and to add comments into the comments section if they wished. When the rater confirmed they were ready to proceed, the next audio was played, and so on until all had been scored.

Phase 3: Stimulated Recall Interviews

After all audios had been scored, raters were given instructions on how to approach the verbal reports. They were advised to listen to each audio again and try to remember and recount what they were thinking when they made their judgements. They were explicitly told to recount what they were thinking during the scoring process, not what they thought at a later time.

Raters recruited through Cambridge Assessment were advised that the researcher did not work for that organisation and that anything they said would not be associated with them in any communications with the organisation in public or private. This was assured in writing as part of the researcher's ethical responsibilities as described in Section 3.4.

At this stage, the rater was told that the audio recorder was being switched on. After raters had confirmed they were happy to proceed they were asked to look at the first rating sheet again and consider the score they had assigned for intelligibility, then the researcher played the audio again. After they had heard the recording, they were asked to report what they had been thinking when they rated that speaker for intelligibility. The prompt used to ask the raters to speak about their rating process follows Leighton & Gierl (2007: 153) as, "Can you tell me all that you can remember thinking as you scored this speaker for intelligibility?"

Raters were not prompted for suprasegmental features. They were informed that the study was about scoring pronunciation but not advised that the focus of the research was on suprasegmental features. This was to avoid the raters telling the researcher what they thought he wanted to hear (Taylor, Bogdan, & DeVault, 2016: 72).

Steps were taken to ensure the validity of the verbal reports. Lyle (2003: 865-6) makes several recommendations to limit the methodological challenges associated with this procedure and these were employed in the following way:

1. Rater anxiety was reduced as much as possible so that they felt less pressure to remodel their responses to please the researcher.
2. The researcher's probes were kept neutral so as to limit the risk of raters perceiving them as judgemental, thus promoting truthful responses. Prompting was undertaken tentatively and raters were consistently advised that they may not, in fact, be able to respond to the researcher's prompts since they were not being asked to generate their response now but to report on memories.
3. The verbal reports were gathered immediately after the rating procedure to increase the probability that the memories would be available for retrieval.
4. The interviews were unstructured so raters were less likely to be led away from their authentic memories.

The procedure for undertaking the stimulated recall was robust and adequate for ensuring raters were reporting their memory of the rating process. It was encouraging to find that some raters were willing to admit they could not remember certain things about the test takers, as it indicates they were willing to only report authentic memories, as described in Chapter 5. However, long-term memory is fallible (Ericsson & Simon, 1985: 19) and raters wilfully misleading the researcher and reporting something other than their original thought processes could not be prevented. Relying on participant truthfulness is a limitation of much qualitative research; however, having a relatively large number of verbal reports per test taker does promote concurrent validity by allowing rater responses to be checked against other sources of data.

3.4 The Pilot

The pilot study referred to throughout this chapter involved gathering scores and verbal reports from 10 participants of whom five had a minimum of three years' experience as raters in English speaking tests and five had no such experience. Ten audio clips which had been selected due to their salient suprasegmental features were presented to the raters who scored them on four 7-point ordered category items for intelligibility, comprehensibility, fluency, and overall pronunciation. Following the scoring, the raters underwent a stimulated recall exercise to recount which features had influenced their scores for intelligibility.

The purpose of the pilot was to gauge the feasibility of the research design. The major findings were that both experienced and non-experienced raters could remember and relate how the suprasegmental features had influenced their decisions. There was a marginal but noticeable difference between the way experienced and non-experienced raters made their decisions in terms of the level of detail and the descriptive terms used. However, this did not impact the consistency of scores which was found to be high, both within the two rater groups and across all raters.

The pilot raised several issues which were resolved in preparation for the main study:

1. Audio clips were increased in time to approximately 60 seconds each (see Section 3.3.1).
2. A standard setting procedure was introduced for the main study (see Section 3.3.1).
3. A context sentence was included on the rating sheet (see Section 3.3.2).
4. 'Overall pronunciation' was removed as a criterion from the rating sheet (see Section 3.3.2).

3.5 Ethical Considerations

Ethical approval was received from the University of Bedfordshire and permission to use the audio clips obtained from Cambridge English Language Assessment.

The purpose of the study was communicated when recruiting participants and again verbally immediately before data collection. The researcher made it clear to raters that the focus of the study was on identifying what influenced their judgements. Informed consent was obtained prior to each rating session by verbally describing the project to the participants. They were then asked to read an informed consent form and sign it if they were willing to proceed. They were encouraged to query anything they did not understand or to raise any concerns they had. Further verbal consent was obtained before the audio recorder was turned on prior to the stimulated recall stage of the data collection.

Anonymity was protected by assigning each rater a unique ID. This code replaced the raters' names on all documents except one, which contained participants' contact details. This was retained so that raters who had requested a report on the outcome of the research could be contacted once the project was completed. It was securely stored on a password protected spreadsheet.

The audio stimuli, survey responses, scores, interview recordings, and transcriptions were all saved on a password protected laptop. The scores and interview responses were also backed up on an encrypted pen drive.

3.6 Summary

This investigation utilised both qualitative and quantitative methods, specifically a rating scale made up of a series of ordered category items, and a stimulated recall procedure. Thirty participants separated into two groups of experienced and non-experienced raters took part in the data collection, assigning scores and reporting their scoring process for 12 test takers which had been selected on the basis of their delivery not conforming to suprasegmental L1 norms. These methods were selected because they are capable of investigating both the outcome of the language testing encounter (i.e. the score) and the rating process underlying that outcome. Score data allows rater performance to be assessed for consistency, and verbal report data reveals rater motivations for assigning their scores which would otherwise be concealed.

Chapter 4: Quantitative Findings and Discussion

The purpose of the quantitative analysis is to answer RQ1 and RQ2 by addressing whether raters are consistent in scoring test takers whose performances are differentiated primarily by suprasegmental characteristics, and by exploring the role of rater experience in judging intelligibility. The analytical approach is discussed in Section 4.1, after which the findings are described in 4.2 and discussed in 4.3.

4.1 Quantitative Analysis

Thirty raters scored 12 test takers on a 9-point intelligibility scale. Each rater judged each test taker so the data matrix was fully crossed. The rating scale (see Appendix C) consisted of a prompt question asking raters how easy it was to recognise the speaker's words and utterances. Each category on the scale was numbered from 1 to 9, with the descriptors "very easy" and "very difficult" at the extreme points. Each rater was assigned an ID prefixed by ER_ if they were in the experienced group or NR_ if they were in the non-experienced group.

Many-Faceted Rasch Measurement (MFRM) analysis was undertaken using the Facets program, version 3.71 (Linacre, 2014), and additional analysis in R, version 3.3.1 (R Core Team, 2016). The facets identified in this analysis were 'rater', 'test taker', and 'rater experience'. The model which specifies how these facets were to be analysed was: $?, ?, ?, R9$. This instructed the Facets program to analyse the interaction between any test taker judged by any rater and in any rater

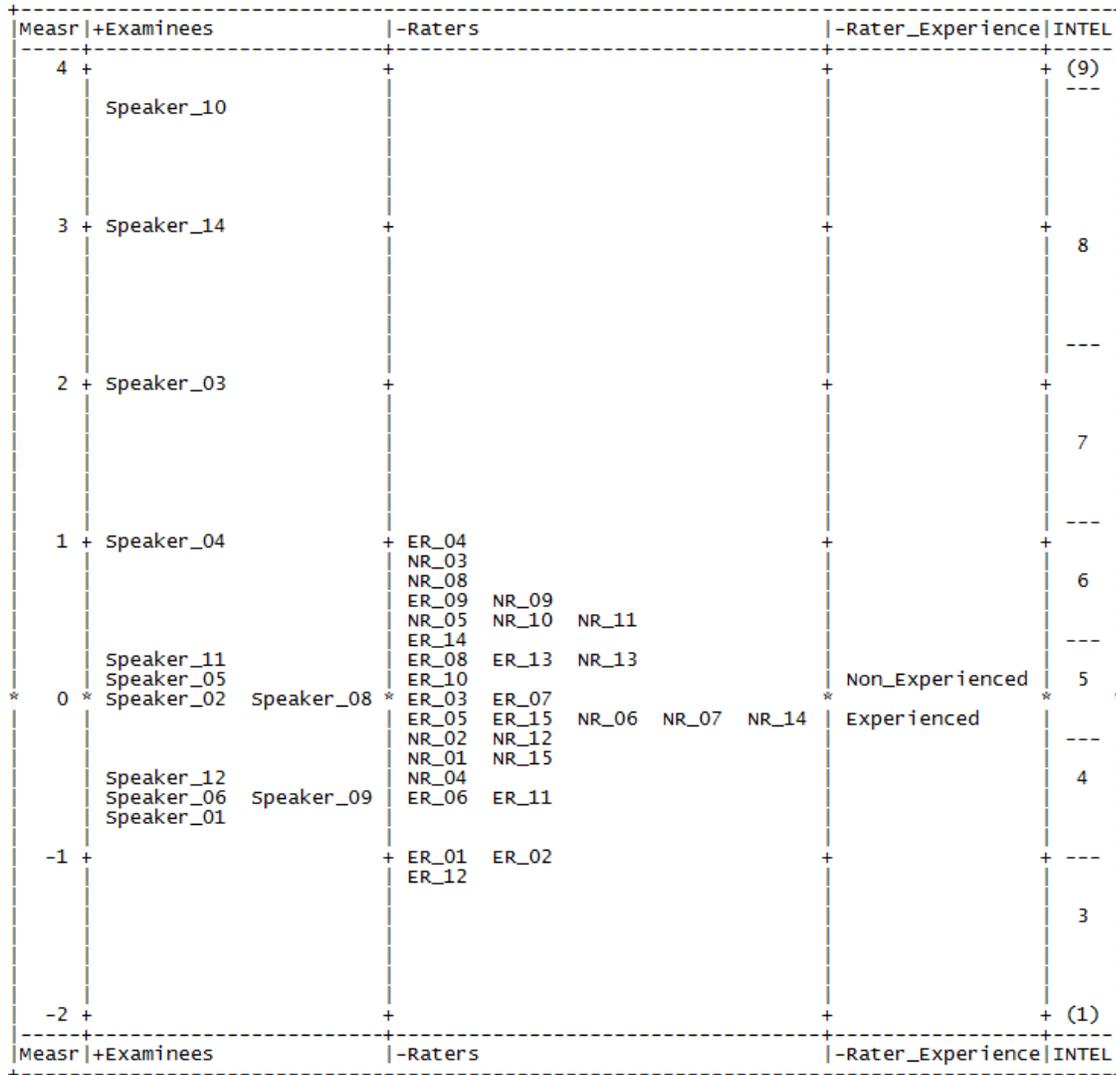
experience group, on a 9-point rating scale. Three analyses were run, once with all raters, once with only the experienced raters, and once with only non-experienced raters. Selected outputs from the Facets analysis are presented at relevant points throughout this chapter and others are reproduced in Appendix H.

4.2 Quantitative Findings

This section begins by reviewing the function of the intelligibility scale. Once the effectiveness of this scale is established, each research question is addressed by examining the consistency of scoring within the experienced rater group, and the difference between the experienced and non-experienced rater groups.

Figure 4.1 below is a variable map summarising the output of the MFRM analysis for all raters. This table shows the intelligibility level of test takers in logits (the 'Measr' column) and the severity of raters and rater groups. Higher logit scores for examinees indicates greater intelligibility. The variable map shows that the test taker graded as most intelligible was Speaker 10 and the least intelligible was Speaker 1. Higher logit scores for raters and rater groups indicate greater severity. The most severe rater was ER_04, the most lenient ER_12. As indicated by the Rater Experience column, the non-experienced raters as a group were more severe than the experienced raters.

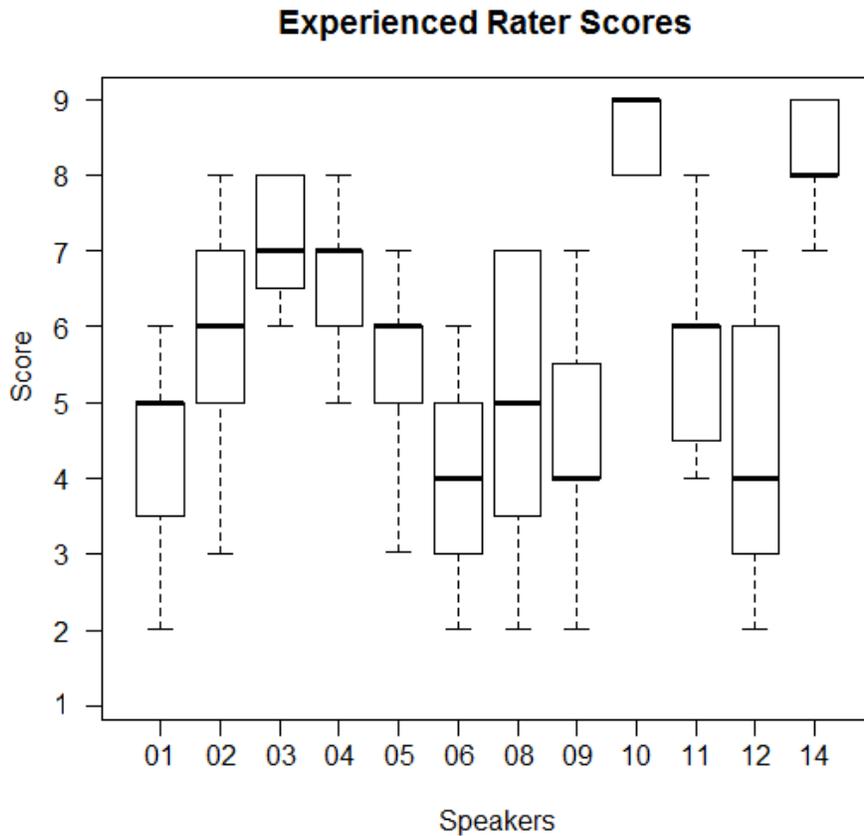
Figure 4.1: Intelligibility Variable Map



Score Distribution

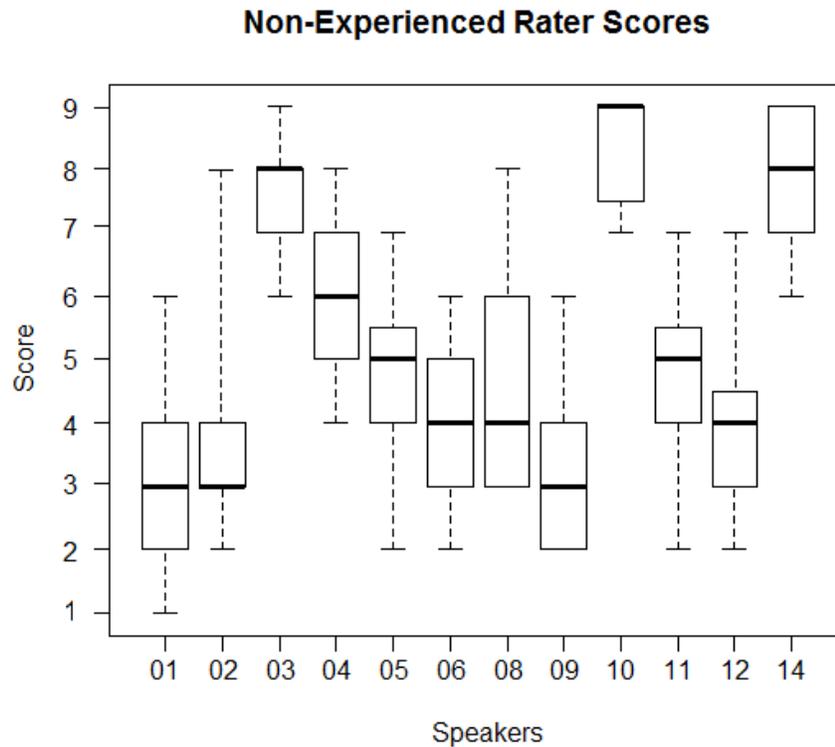
The box plots below (Figures 4.2, and 4.3) show the raw score distributions for each group of raters. For each speaker, the thick central bar represents the median score they received, the segments above and below the median represent the first and third quartiles, and the whiskers above and below the quantile segments represent the minimum and maximum values. Raw scores are reported in Appendix I.

Figure 4.2: Experienced Rater Scores by Audio Clip



The box plot for experienced raters shows some variation in rater scores. Several speakers were assigned scores within a narrow band, such as Speaker 3 who received scores of 6 to 8, and for others there was more variation such as for Speaker 8 who received scores from 2 to 7. Raters avoided the lowest point on the scale and there is evidence to suggest there might have been a ceiling effect, whereby rater scores were limited at the top end due to the scale being too short. According to Fisher (2007) a ceiling effect may be present if more than 5% of scores fall into the highest category. In the current study 8% of scores fall into that category.

Figure 4.3: Non-Experienced Rater Scores by Audio Clip



The box plot of non-experienced rater scores shows a similar degree of variation. The lowest point on the scale was also not well utilised, it was only used once. There is also evidence of a possible ceiling effect for Speakers 10 and 14, and for non-experienced raters a top score of 9 was also assigned to Speaker 3.

Scale Function

Despite the potential for a ceiling effect, the output of the facets analysis indicates that the points are categorically functional. The separation value of all test takers was 8.25, as illustrated in Table 4.1 below. This indicates that all raters separated the test takers into approximately eight separate ability levels. However, when raters are analysed in their separate groups, this drops to 6.78 for experienced raters and 5.4 for non-experienced raters. Therefore, experienced raters appear to

be separating the test takers into a greater number of distinct levels of intelligibility than the non-experienced raters. This may be due to experienced raters having a more nuanced understanding of intelligibility, and therefore being able to discern greater degrees of difference among the speakers. Alternatively, it may be caused by non-experienced raters being more tentative about using scales and averse to using the extreme points.

Table 4.1: Test Taker Summary Statistics

Statistics	Test takers (All Raters)	Test takers (Exp Raters)	Test takers (Non-Exp Raters)
Mean Measure (Logits)	0.64	0.36	0.38
Mean Standard Error	0.18	0.29	0.24
Chi-square ($p < 0.01$)	521.8	322	255
d.f.	11	11	11
Separation Index	8.25	6.78	5.4
Separation Reliability	0.99	0.98	0.96

Table 4.2 below shows average intelligibility estimates increasing as the scale points increase. This suggests that raters agree that the scale represents a progression from low to high intelligibility, indicating that raters are using the scale in a consistent manner. Note, though, that the intelligibility estimates for category one may not be valid given that it was only assigned once, and fewer than ten observations per category can result in the step calibration estimates being imprecise or unstable (Linacre, 1999: 108). Point 1 on the scale was never assigned by an experienced rater, which explains why there is no intelligibility estimate at Category 1 for those raters.

Table 4.2: Average Intelligibility Estimates

Category (score)	Total	Average Intelligibility Estimates (Logits)		
		All Raters	Experienced Raters	Non-Experienced Raters
1	1	-1.39		-1.4
2	20	-1.09	-2.4	-1.02
3	48	-0.68	-1.71	-0.6
4	64	-0.42	-1.33	-0.4
5	46	0.03	-0.62	0.09
6	57	0.58	-0.03	0.51
7	52	1.41	0.77	1.41
8	42	2.30	2.59	1.83
9	30	3.61	4.8	2.79

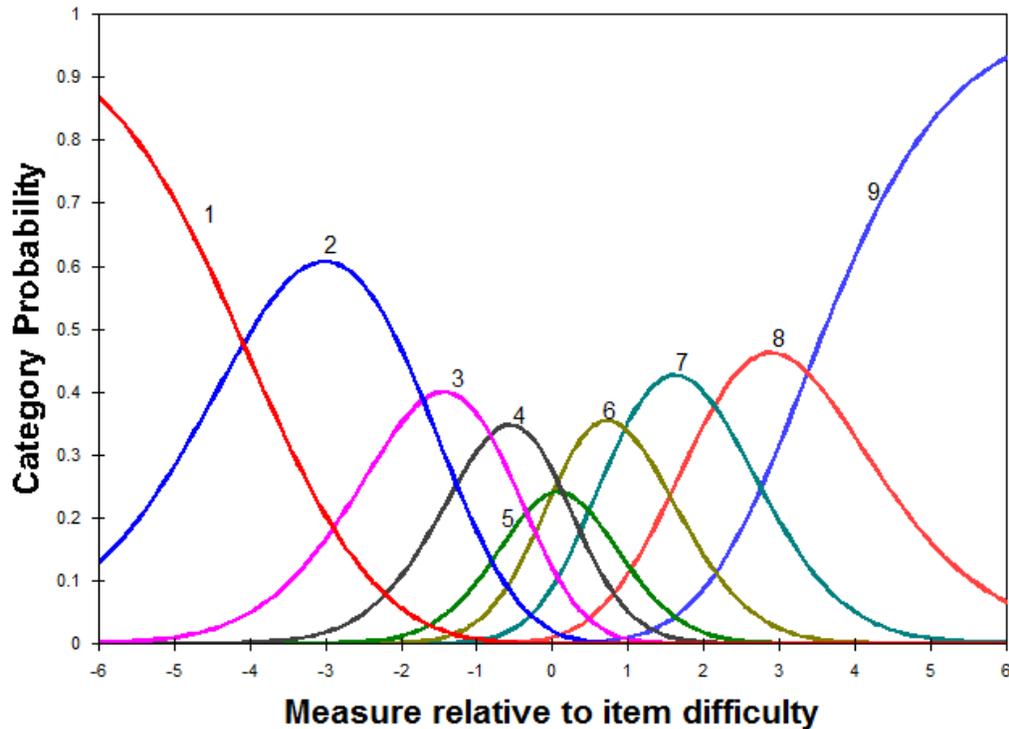
As described in the rationale for using MFRM analysis (Chapter 3), the model is probabilistic, using the pattern of scores to predict how raters of a given severity will respond to test takers of a given ability. In the rating scale model, the Facets program translates this into a value which represents the point at which adjacent categories on the scale are equally likely to be selected. This is known as the Rasch-Andrich Threshold. Table 4.3 reports the Rasch-Andrich Thresholds for the intelligibility rating scale. It shows that when all raters are analysed together, a test taker with ability of -1.71 logits was equally likely to be assigned a score of 2 as 3, for instance, and at -0.82 logits they are equally likely to be assigned 3 as 4.

Table 4.3: Rasch-Andrich Thresholds

Category (score)	Total	Rasch-Andrich Thresholds (Logits)		
		All Raters	Experienced Raters	Non-Experienced Raters
1	1			
2	20	-4.08		-3.69
3	48	-1.71	-2.94	-1.59
4	64	-0.82	-1.92	-0.70
5	46	0.13	-0.98	0.41
6	57	0.01	-0.84	0.27
7	52	0.96	0.53	0.73
8	42	2.14	1.80	2.05
9	30	3.37	4.35	2.51

Despite the increase in average intelligibility estimates shown in Table 4.2 above, there is disorder in the Rasch-Andrich Thresholds for all raters, which increases to 0.13 until category five, and then falls to 0.01 in category six before increasing again. When each rater group is analysed separately, the disorder disappears for experienced raters but remains in place for the non-experienced raters. The disorder can be noted graphically in Figure 4.4 below, which shows the probability curves for each category and all raters. These curves represent the probability of each category on the scale being observed at a given logit score. So, at -6 logits the probability of raters assigning a score of 1 approaches 0.9 and the probability of a 2 is between 0.1 and 0.2. The Rasch-Andrich Thresholds are the points at which two category curves cross. The peak of each curve represents the point at which a particular category is most likely to be selected. For example, a score of 7 is most likely to be assigned for a candidate with a logit ability of approximately 1.75.

Figure 4.4: Scale Category Probability for All Raters



In Figure 4.4 all categories have a distinct peak, except for Category 5. In other words, for all raters as a group and the non-experienced rater group there are intelligibility levels at which each category on the rating scale is the most likely to be selected apart from Category 5, which is never the most likely to be selected. The probable reason for this disorder is the lower number of observations in Category 5 relative to adjacent categories, suggesting that non-experienced raters considered point 5 on the scale to be a narrower intelligibility interval than the others. Note that non-experienced raters did have monotonic progression of intelligibility, as demonstrated above in Table 4.2, so they appear to recognise the progression of ability among the sample of speakers, but perhaps they are unsure how to apply point 5. As a result of this disorder, collapsing Category 5 into another category was considered (as advised by Salzberger, 2015), but since the disorder is with the midpoint it is not clear which category it should be merged with (i.e. point 4 or

point 6). Nonetheless, if this scale was going to be used again by non-experienced raters it would be prudent to examine why they avoided using point 5 and to take appropriate action.

Model Fit

The final way in which the scale points can be investigated is by reviewing the fit indices. Fit indices report the difference between observed and expected scores. In other words, they relate the difference between the raw scores assigned by raters and the scores expected by the assumptions of the Rasch model. Two fit statistics are reported: infit and outfit. Infit is information weighted and gives more value to on-target observations, as opposed to outfit, which includes unexpected outlying observations. Both fit statistics are reported as mean-square values ranging from 0 to infinity, where values <1 indicate muting or scoring patterns that are too predictable, and where values >1 indicate noise or too much randomness (Linacre, 2013).

In terms of how well the scale points fit the expectations of the Rasch model, Linacre (1999) recommends that outfit mean-square values greater than 2 distort measurement by incorporating too much randomness. Table 4.4 below reports outfit mean-square values for the categories on the intelligibility scale. These values range from 0.7 to 1.3 for experienced raters and from 0.5 to 1.7 for non-experienced raters. This indicates that non-experienced raters' selection of scale points fits less well than that of the experienced raters for some of the speakers. Overall, however, none of the values are over 2, suggesting that the data fits the Rasch model sufficiently well and does not incorporate too much randomness.

Table 4.4: Category Outfit Mean-Square

Category (score)	Total	Outfit Mean Square		
		All Raters	Experienced Raters	Non-Experienced Raters
1	1	0.9		0.9
2	20	0.9	0.7	0.9
3	48	0.9	0.9	1
4	64	0.8	0.7	1
5	46	0.8	0.9	0.5
6	57	1	1	1.2
7	52	1.1	1.3	0.9
8	42	1.1	1	1.7
9	30	0.8	0.9	0.7

Infit is particularly useful when judging individual rater performance because it measures how closely raters fit the Rasch model without being overly influenced by outlying judgements. Bond and Fox (2015: 273) recommend that individual rater infit mean-square values <0.6 indicate too much predictability and values >1.4 too much randomness. Predictability does not undermine rater judgements because muted scores do not detract from the measurement, even if they do not particularly contribute to it either (Linacre, 2002). By contrast, randomness is considered to add misinformation to the measurement and is therefore relevant to the assessment of rater performance. Although Linacre recommends a range of 0.5 - 1.5 for mean-square statistics, he suggests that it is only values >2 that distort the measurement (Linacre, 2002).

When the rater groups are analysed separately no raters distorted the measurement by having a misfit value greater than 2. However, several raters breach the recommended guidelines of 0.5 – 1.5. Table 4.5 below shows rater infit mean-square values by rater group. Experienced raters ranged from 0.44 – 1.72, and non-experienced raters 0.25 – 1.95. This suggests that, as a group, the non-experienced raters exhibited somewhat more randomness and predictability than experienced raters. Two experienced raters and two non-experienced raters tended to

predictability (infit <0.5), whereas one experienced rater and two non-experienced raters tended to randomness (infit >1.5).

Table 4.5: Rater Performance by Group

Experienced Raters		Non-Experienced Raters	
Rater	Infit Mean Square	Rater	Infit Mean Square
ER_08	1.72	NR_15	1.95
ER_06	1.45	NR_02	1.86
ER_13	1.32	NR_11	1.38
ER_01	1.3	NR_04	1.25
ER_09	1.21	NR_06	1.22
ER_14	0.99	NR_13	1.15
ER_05	0.95	NR_12	1.04
ER_15	0.88	NR_05	0.98
ER_11	0.76	NR_07	0.81
ER_12	0.76	NR_10	0.7
ER_07	0.7	NR_08	0.54
ER_02	0.62	NR_01	0.54
ER_10	0.6	NR_03	0.5
ER_04	0.45	NR_14	0.35
ER_03	0.44	NR_09	0.25

Table 4.6 reports individual rater performance when all raters are analysed together. In this case, one rater had an infit mean-square >2, NR_15 (infit: 2.16). Three other raters had infit >1.5 and <2. In this combined analysis, all four raters who had a tendency for randomness were non-experienced and half of the six raters who tended towards predictability were experienced raters.

Table 4.6: Rater Fit Statistics (All Raters)

Rater	Infit Mean Square	Rater	Infit Mean Square
NR_15	2.16	NR_07	0.8
NR_04	1.93	ER_15	0.77
NR_02	1.79	ER_05	0.75
NR_06	1.57	ER_07	0.61
NR_11	1.43	NR_03	0.58
NR_05	1.38	ER_11	0.54
NR_13	1.31	NR_08	0.53
ER_13	1.3	ER_12	0.51
ER_06	1.27	ER_03	0.51
ER_09	1.22	ER_04	0.47
ER_08	1.15	NR_01	0.45
NR_12	1.11	NR_14	0.43
ER_14	1.03	ER_02	0.39
ER_01	0.99	ER_10	0.37
NR_10	0.89	NR_09	0.25

Exactly how the misfitting rater deviated when all the raters are analysed together is not obvious from observing the raw scores. Identifying patterns in how raters deviated was approached systematically by calculating z-scores for each category and each rater:

$$z \text{ score} = \frac{N - \text{mean}(N)}{\text{standard deviation}(N)}$$

N = Number of times a category was selected

This standardises the scores by referring to them in terms of how many standard deviations they are from the mean. These z-scores are presented in Appendix G. Seven ratings stand out as being more than two standard deviations from the mean (>1.96 or <-1.96). The z-scores of the misfitting raters shows that NR_15 assigned a top score rather more commonly than other raters (25% of the time), but they do not uncover any errant behaviour on the part of the borderline raters NR_02, _04 and _06. This suggests that the misfit exhibited by NR_15 may be due to irregular

leniency, although the borderline misfit attributed to other raters cannot be accounted for by a clearly observed pattern of errant scoring. The finding that, when analysed separately, no raters had infit greater than 2 indicates that, although some rater's scores were unproductive, they were not degrading (Wright & Linacre, 1994).

Rater Severity

Table 4.7 below shows each rater's fair average score when all raters are analysed together. The fair average score is the average score accounting for individual variation. It is obtained by taking the mean of all scores assigned by a rater and modifying it by their severity estimate.

There was a range of severity. The most severe rater was ER_04, whose fair average score assigned was 4.20 (+1.06 logits), and the least severe rater was ER_12, whose fair average score assigned was 7.14 (-1.13 logits). This amounts to a difference of 2.72 scale points. The range of severity of experienced raters when analysed separately was the same at 2.72 (4.56 to 7.28), and for non-experienced raters it was smaller at 2.36 (4.01 to 6.37). The Facets outputs pertaining to these figures are reproduced in Appendix H.

Table 4.7: Rater Performance (All Raters)

Rater	Fair Average Score	Severity Estimate (Logits)
ER_12	7.14	-1.13
ER_01	6.99	-0.97
ER_02	6.99	-0.97
ER_11	6.69	-0.66
ER_06	6.61	-0.59
NR_04	6.55	-0.53
NR_01	6.4	-0.4
NR_15	6.32	-0.34
NR_02	6.16	-0.2
NR_12	6.16	-0.2
ER_15	6.13	-0.18
NR_06	6.07	-0.14
NR_07	6.07	-0.14
ER_05	6.04	-0.12
NR_14	5.98	-0.07

Rater	Fair Average Score	Severity Estimate (Logits)
ER_07	5.95	-0.05
ER_03	5.86	0.01
ER_10	5.67	0.14
ER_13	5.57	0.21
ER_08	5.57	0.21
NR_13	5.48	0.26
ER_14	5.36	0.34
NR_05	5.14	0.48
NR_10	5.14	0.48
NR_11	5.02	0.55
ER_09	4.91	0.62
NR_09	4.9	0.62
NR_08	4.65	0.78
NR_03	4.39	0.94
ER_04	4.2	1.06

The Facets software provides a reliability index which reflects the reliability of these differences in rater severity (i.e. whether they can be accounted for by estimate error). A high reliability index suggests there is a high probability that the difference in severity values reflects real differences in rater severity (McNamara, 1996: 140). Table 4.8 shows that for raters in the current study the reliability index was 0.77 for all raters, 0.86 for experienced raters, and 0.66 for non-experienced raters, indicating that the differences in rater severity can be relied upon to be accurate.

Table 4.8: Rater Summary Statistics

Statistics	All Raters	Experienced Raters	Non-Experienced Raters
Mean Measure (logits)	0.00	0.00	0.00
Mean Standard Error	0.27	0.3	0.26
Chi-square ($p < 0.01$)	117.7	90	40.1
d.f.	29	14	14
Separation Index	1.85	2.46	1.4
Separation Reliability	0.77	0.86	0.66

The raters appear to make their judgements approximately in line with the judgements made by Cambridge Examiners. Table 4.9 shows the test takers ordered by exam and then Cambridge Assessment pronunciation scores. It shows that intelligibility judgements approximately follow speakers as they proceed through exam levels and Cambridge Assessment pronunciation scores. Note though, that for Speakers 1 and 5 pronunciation scores were unavailable.

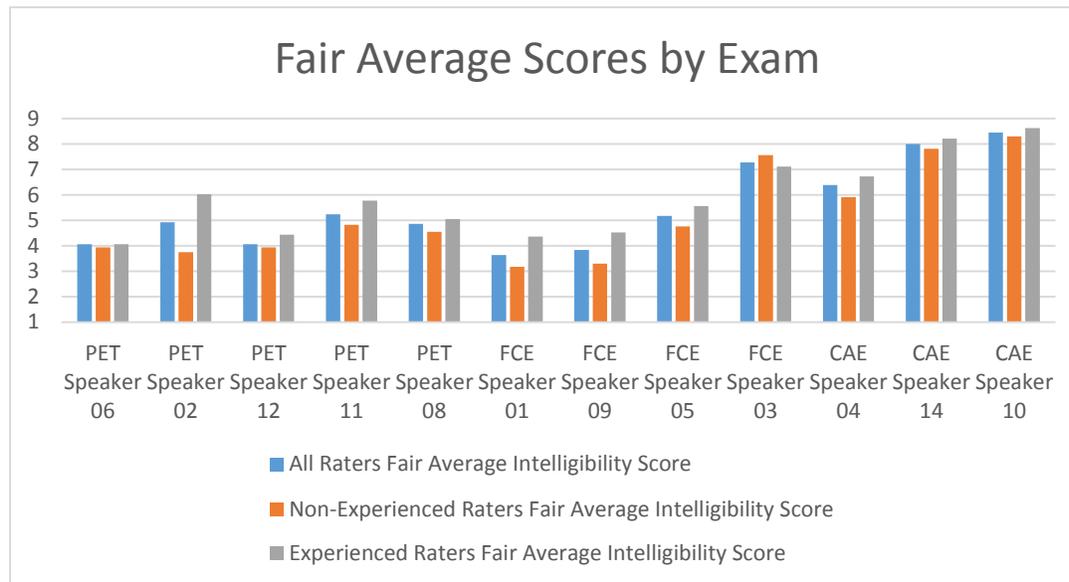
Table 4.9: Test Taker Fair Average Scores by Exam

Test Taker	Exam	Cambridge Assessment Pronunciation Score	All Raters Fair Average Intelligibility Score	Non-Experienced Raters Fair Average Intelligibility Score	Experienced Raters Fair Average Intelligibility Score
Speaker 06	PET	2.5	4.06	3.94	4.06
Speaker 02	PET	3	4.93	3.75	6.03
Speaker 12	PET	3.5	4.06	3.94	4.44
Speaker 11	PET	3.5	5.24	4.83	5.77
Speaker 08	PET	4	4.86	4.55	5.05
Speaker 01	FCE	unavailable	3.63	3.17	4.36
Speaker 09	FCE	2.5	3.84	3.3	4.52
Speaker 05	FCE	unavailable	5.17	4.76	5.56
Speaker 03	FCE	3.5	7.28	7.56	7.12
Speaker 04	CAE	4	6.39	5.91	6.73
Speaker 14	CAE	4	8	7.82	8.22
Speaker 10	CAE	4	8.45	8.3	8.63

Figure 4.5 below illustrates this graphically. The test takers are listed across the bottom of the chart progressing from lowest level exam (PET) to highest level exam (CAE), and ordered by Cambridge Assessment score within each group. There is a clear distinction at the right side of the figure, with the CAE candidates and one of the FCE candidates appearing to receive higher fair average scores by all test takers. There is a less clear distinction between the PET and lower-level FCE candidates. This may be accounted for by the way these exams cross over at certain

CEFR levels. For instance, PET assesses candidates from A2 to the lower half of B2 on the CEFR, whereas FCE covers B1 to the lower half of C1 (Cambridge English, 2015: 3). Furthermore, the Cambridge Assessment pronunciation scores are based on a range of criteria rather than simply global intelligibility, as can be observed in the Cambridge English Overall Speaking Scale presented in Appendix B.

Figure 4.5: Fair Average Scores by Exam



Rater Reliability

Despite the variation in severity, raters were found to be reliable according to a point-biserial correlation index (the Single Rater-Rest of Rater Correlation column in Table 4.10 below). Within each group, raters ranged from 0.66 to 0.96, and as a group on average the non-experienced raters were slightly more consistent with a mean of 0.85 versus 0.82 for the experienced raters. This demonstrates that there was agreement among raters within each group concerning the ordering of the performances. There were no values lower than 0.3, which is the level below which ratings are considered to be inconsistent (Prieto & Nieto, 2014: 388).

Table 4.10: Single Rater-Rest of Raters Correlation within Groups

Experienced Raters	SR-RR Correlation (All Raters)	SR-RR Correlation (Exp Raters)	Non-Experienced Raters	SR-RR Correlation (All Raters)	SR-RR Correlation (Non-Exp Raters)
ER_01	0.7	0.7	NR_01	0.93	0.91
ER_02	0.83	0.81	NR_02	0.71	0.7
ER_03	0.9	0.9	NR_03	0.94	0.95
ER_04	0.95	0.94	NR_04	0.72	0.79
ER_05	0.79	0.8	NR_05	0.84	0.9
ER_06	0.64	0.66	NR_06	0.82	0.85
ER_07	0.8	0.81	NR_07	0.84	0.82
ER_08	0.82	0.78	NR_08	0.92	0.92
ER_09	0.77	0.8	NR_09	0.95	0.96
ER_10	0.93	0.9	NR_10	0.8	0.86
ER_11	0.87	0.86	NR_11	0.89	0.9
ER_12	0.81	0.81	NR_12	0.64	0.67
ER_13	0.85	0.86	NR_13	0.8	0.83
ER_14	0.76	0.8	NR_14	0.89	0.91
ER_15	0.85	0.87	NR_15	0.74	0.75
Mean	0.82	0.82	Mean	0.83	0.85
StDev	0.08	0.07	StDev	0.09	0.09
Min	0.64	0.66	Min	0.64	0.67
Max	0.95	0.94	Max	0.95	0.96

This finding is supported by the arithmetic mean of weighted Cohen's Kappa for all pairs of raters, which was 0.56 for experienced raters, 0.62 for non-experienced raters, and 0.57 for all raters. This represents moderate agreement among raters and supports the conclusions of the facets analysis that raters broadly agree on scores. It also reasserts the finding that non-experienced raters are marginally more consistent than the experienced raters.

Group Severity

When all raters were analysed together the facets program provided a comparison of rater groups. Table 4.11 shows the difference between experienced and non-experienced rater groups. The fair average score of the experienced rater group was 6.07, while the non-experienced rater

group's fair average score was 5.67. This difference amounts to 0.4 scale points by fair average. Testing the hypothesis that all rater groups are equally severe with a fixed chi-square test demonstrates that this effect is not simply due to rater error ($\chi^2 = 8.7$, $df = 1$, $p < 0.01$).

Table 4.11: Rater Group Severity

Rater Group	Total score	Count	Observed average	Fair Average Score	Intelligibility Measure (Logits)	SE	Infit Mean-Square	Outfit Mean-Square
Experienced	1052	180	5.84	6.07	-0.14	0.07	0.8	0.81
Non-Experienced	931	180	5.17	5.67	0.14	0.07	1.12	1.11
Mean Measure (logits): 0.00, Mean Standard Error: 0.07, Chi-square ($p < 0.01$): 8.7, d.f.: 1, Separation Index: 2.77, Separation Reliability: 0.88								

Summary of Quantitative Findings

1. Experienced raters assigned scores consistently and fit the Rasch model adequately. They appeared to use the scale effectively despite evidence for a ceiling effect on some speakers.
2. Non-experienced raters also assigned scores consistently and fit the Rasch model adequately. Broadly they used the scale effectively, although they exhibited some Rasch-Andrich disorder around the mid-point and a possible ceiling effect on some speakers. When all raters were combined into a single analysis, one of the non-experienced raters was found to be excessively random in scoring.
3. Non-experienced raters exhibited a similar range of severity to experienced raters, although they were found to be somewhat more severe and slightly more consistent with one another in their scores than experienced raters.

4.3 Quantitative Discussion

This section concerns answering RQ1 and RQ2. However, prior to addressing these questions it is necessary to discuss the extent to which the intelligibility scale functioned effectively. The analysis has succeeded in demonstrating that a 9-point intelligibility scale works effectively. Intelligibility estimates were found to progress monotonically. This suggests that raters broadly understood how to use the scale and applied it in a way consistent with intelligibility being a measurable trait. Non-experienced raters seemed to have had some difficulty applying the mid-point of the scale, as that point was assigned fewer times than the adjacent ones. Isaacs and Thomson (2013) found that scale category overlap was present in similar scales with both 5 and 9 points, so the disorder found around the central point in the intelligibility scale was not unexpected. It is unclear from the qualitative or quantitative elements of this study what caused the non-experienced raters to avoid the mid-point, however it may be consistent with the findings of Isaacs and Thomson (2013), i.e. due to raters being “unable to meaningfully differentiate between ‘so many numbers’” (p154). Given that the experienced raters did not appear to exhibit this phenomenon it may be the case that understanding the use of a mid-point on a scale is something that emerges from training or experience of using such scales. In short, it appears to be related to a rater’s ability to use the scale rather than any perceptual inability.

There was evidence of a ceiling effect in both rater groups, perhaps caused by the fact that the speakers were already rather intelligible (minimum CEFR B1) and could be expected to appear relatively high on the scale. Other possibilities are rater reluctance to appear overly critical (Tourangeau, Rips, & Rasinski, 2000), although the qualitative analysis presented in Chapter 5 provides no suggestion of this. The impact of the ceiling effect is that it may be muting the scores of the three highest scoring test takers. However, it does not affect the ordering of test takers in terms of intelligibility, and so, for the purposes of this project, it does not represent a major risk to

the interpretation of scores. Finally, the fit statistics showed that although some of the raters were borderline in terms of the predictability and randomness of their judgements, none were found to breach the rule that Linacre (2002) sets out of *infit* being greater than 2 when the groups were analysed separately. In short, it appears that experienced and non-experienced raters could accurately and reliably use a 9-point scale to measure intelligibility. RQ1 and RQ2 will now be discussed.

4.3.1 Consistency within the Experienced Rater Group [RQ1]

RQ1 sought to identify whether experienced raters were consistent in the way they scored performances that were differentiated primarily on basis of the speaker's use of non-standard suprasegmental features. The findings were that experienced raters were capable of consistently judging the intelligibility of such speech samples. Raters made consistent judgements across a range of suprasegmental variables, specifically relating to lexical stress, rhythm, and intonation. In other words, despite each having their own unique listening experience and rating background, individual raters appear to be able to consistently score samples of speech which are marked by non-standard suprasegmental features of speech. These errors therefore do not appear to provoke severe inconsistency in scores.

Inconsistent scoring challenges validity fundamentally because without scores being equivalent among different raters, it is difficult for their meaning to be interpreted. It is not possible to tell what a speaker's true level of proficiency is, or indeed whether a true level of proficiency exists. The purpose of exploring rater consistency of scores, therefore, was based on consideration for reliability, which facilitates meaningfulness and hence results in validity.

Consistency is often found in laboratory studies that examine intelligibility using, for instance, transcription tasks. It is also a feature of studies where raters have judged intelligibility on a scale,

although it is not always clear how intelligibility is defined in such studies (e.g. Llorca, 2000; Lumley & McNamara, 1995). It was hypothesised in Chapter 2 that experienced raters would score intelligibility consistently. Existing literature shows that criteria such as accent, comprehensibility and fluency can be judged on a 9-point scale (e.g. Munro & Derwing, 1995a, Derwing & Munro, 1997), so it was also predicted that raters would be adequately capable of using a 9-point intelligibility scale. The findings here have confirmed this hypothesis.

Oral proficiency studies measure a range of constructs, such as accentedness, comprehensibility, and fluency, but the narrow definition of intelligibility as perceptual recognition of words and utterances has rarely been measured on a rating scale in reference to language assessment. For this reason it is difficult to corroborate the findings with other studies in the field. By focusing on suprasegmental features in reference to intelligibility as perceptual recognition, and by grounding the findings unequivocally in the language testing context, the findings presented here extend those of other studies which demonstrate that listeners are consistently influenced by suprasegmental characteristics of the speech signal (e.g. Kang, 2013b; Pickering, 2001; Hahn, 2004; Iwashita et al., 2008).

There are two primary implications to these findings. Firstly, judgements of language proficiency, and especially something as ephemeral as pronunciation, are necessarily somewhat subjective. This means different raters with different backgrounds may make different judgements. It is especially the case for intelligibility where listener interpretation is regarded as part of the construct (Smith & Nelson, 1985: 333). The role of the listener in constructing intelligibility provides theoretical justification for inconsistency of scores, and some researchers go as far as to suggest that inconsistency is desirable, at least initially where it reflects a range of perspectives and expertise (Shohamy 1994: 99; 1995: 206). Furthermore, it was not clear that all suprasegmental features are interpreted by listeners in a consistent way. Chapter 2 demonstrated

that there is evidence of certain features having a consistent influence on certain pronunciation proficiency criteria; however, rater perceptual recognition of speech signals that contain suprasegmental deviations from norms is not well understood. The first implication of this finding is, therefore, that it demonstrates for the first time that raters are able to make consistent judgements on a 9-point intelligibility scale, that is, a scale focusing on perceptual recognition in contrast to ease of understanding. By targeting rater perception in this way, this project has provided new insight into how raters attend to the initial level of processing of non-native speech, and indicates that intelligibility is a suitable criterion for use in measuring pronunciation.

The second implication is that it begins to address the issue of whether suprasegmental features of speech are suitable to be included in rating scales by demonstrating that listener perceptual recognition of suprasegmentally errant deliveries is consistent. Without such a common understanding among raters, the inclusion of suprasegmental descriptors on rating scales may have a negative impact on reproducibility of results. This, in turn, may influence the ability of score users to discern the meaning of those scores. However, it must be noted that the quantitative findings present no direct evidence of the influence of suprasegmental features of speech on scores. Raters might have been wholly ignoring the suprasegmental element of the performance and attending only to the segmental characteristics. Even with audio clips that primarily target suprasegmental features, questions remain over how influential specific features were, and whether the suprasegmental features identified by the phoneticians can be expressed using terms that would be mutually comprehensible across raters. This is addressed using qualitative data in the following chapter.

4.3.2 Influence of Rating Experience [RQ2]

RQ2 required a between-groups analysis of experienced and non-experienced raters. Three levels of consistency were addressed: 1) the difference in rater reliability between each group; 2) the difference in the range of scores assigned by each group; and 3) the difference in severity between each group.

In Chapter 2 it was hypothesised that experienced raters would find speakers rather more intelligible due to their linguistic experience. Multiple trace theory was employed to describe a situation in which linguistically experienced listeners would retain more exemplars of non-native speech in their long-term memory from which to match a test taker's speech signal. In addition to this, it was anticipated that experienced raters would be more reliable in their scores than non-experienced raters. This hypothesis was based on existing studies suggesting that experienced raters are more reliable when judging performance on a scale (e.g Thompson, 1991; Saito et al., 2016). This may be due to the rating procedure being completely new to non-experienced raters but well established in the minds of experienced raters. Finally, it was suggested that both rater groups would have a broad score range (as found in similar such studies, e.g. Davis, 2016; Eckes, 2011; McNamara, 1996).

The results show that the hypotheses were borne out in terms of relative severity by group, and in terms of score range. It was indeed the case that experienced raters found test takers more intelligible than non-experienced raters did; the effect was relatively small, however, amounting to less than half a scale point on the 9-point intelligibility scale. By contrast, the hypotheses were not borne out in reference to rater reliability. Non-experienced raters were found to be slightly more reliable than experienced raters, and they also had a slightly narrower range of scores. These differences are now discussed.

Reliability

Reliability was measured in two ways to avoid a single measure being misleading (Stemler & Tsai, 2008): a point-biserial correlation index, which was an average of 0.82 (standard deviation 0.08) for experienced raters and 0.83 (standard deviation 0.09) for non-experienced raters; and mean weighted Cohen's Kappa for all pairs of raters, which was 0.56 for experienced raters and 0.62 for non-experienced raters.

Both groups of raters were found to be suitably reliable. This conforms to several studies demonstrating that non-experienced raters perform similarly to experienced raters (Isaacs & Thomson, 2013; Bongaerts et al., 1997), and that agreement has been found to be adequate prior to rater training (Davis, 2016; Derwing et al., 2004; Bridgeman et al., 2012). The findings presented here extend such studies by demonstrating that such a phenomenon is also present in judgements of perceptual recognition. A rather counterintuitive finding was that experienced raters were less consistent than non-experienced raters on both measures, although the difference in the point-biserial correlation index was only 0.01. This finding contradicts studies suggesting that experience bestows reliability in judging oral proficiency (Thompson, 1991; Saito et al., 2016; Kim, 2015), although, as discussed in the literature review, 'experience' is by no means a fixed characteristic throughout these studies.

One reason this finding is rather surprising is that the experienced raters had all undergone training to promote consistency in their scoring and the non-experienced raters had not. It seems intuitively sensible that the raters who had received training would be more consistent than those who had not. There has been much research demonstrating that rater training improves rater reliability (e.g. Davis, 2016; Kim, 2015) but limited investigation into whether raters trained to examine a range of different exams score differently on a generic scale, such as the one used in this study. Considering that training typically seeks to standardise raters to a specific analytical

scale, it may be the case that a group of raters who have been trained for a variety of exams would be slightly less consistent than those who have received no training. This could be due to experienced raters internalising the constructs of the various exams they are trained to examine (Harding 2016: 31).

Ten raters had been trained by Cambridge Assessment to assess the speaking component of the Main Suite exams, but raters had also been trained to examine IELTS, Skills for Life, and TOEFL. Therefore, if experienced raters engaged a combination of their previous training when making decisions, they might have been approaching the task from different perspectives. The non-experienced raters had received no training, meaning they had more in common with each other in terms of training than the experienced raters had with each other. On the basis of these results, it does not appear to be the case that having experience of rating results in a greater common interpretation of the intelligibility of non-standard suprasegmental deliveries than typical listeners.

The finding that non-experienced raters were somewhat more consistent than experienced raters has another implication. It suggests that people who do not have rater training and linguistic experience also interpret speech that consists of suprasegmental errors in a common way. Suprasegmental variation, therefore, appears to feature in real-world judgements of intelligibility and does not provoke excessive variation in judgements. This is the case across lexical stress, rhythm, and intonation. Field (2011) argues that “intelligibility to the assessor does not guarantee intelligibility to the wider audience” (p85). These findings indicate that a common perception of the intelligibility of suprasegmentally errant performances is not something that arises through experience or training, but rather is inherent to native listeners. As such, it widens the validity argument to the broader communication context by supporting an extrapolation inferential link (McNamara & Roever, 2006, cited in Harding, 2017: 4) between the exam context and real-world interpretation made by typical listeners.

Score Range

The range in fair average scores among all raters was 2.72 (2.19 logits). This range of severity is consistent with language tests and was not unexpected (Davis, 2016; Eckes, 2011; McNamara, 1996). When separated into rater groups the range of fair average scores assigned by experienced raters was 2.72, and 2.36 by non-experienced raters. This suggests that experienced raters have a somewhat broader range of severity.

Variation in rater severity has been found to be resistant to rater training (Barrett, 2001; Lumley & McNamara, 1995; Weigle, 1998), and it is almost taken for granted in language assessment research, with McNamara regarding it as “a fact of life” (McNamara, 1996: 127). This finding indicates that the issue of rater variation associated with scales generally is also applicable to an intelligibility scale. However, it is somewhat surprising that experienced raters had a broader range than non-experienced raters. A possible explanation for this is central tendency bias on the part of non-experienced raters. They might have been more tentative in their approach, avoiding the extreme points of the scale more than the experienced raters and therefore having a greater degree of consistency. There is some evidence supporting this, as the non-experienced raters' score range was narrower than the experienced raters', indicating that they were less willing to apply a broad range of scores. But the Rasch IRT analysis demonstrated that the non-experienced rater group consisted of raters who exhibited relatively more randomness and more predictability than the experienced rater group. They do not appear to have been using the scale in a more tentative manner than the experienced raters. Instead, it is probable that the source of the difference is that non-experienced raters had more in common with one another as a group than the experienced raters did due to experienced raters' exposure to a range of exam scales and training procedures, and non-experienced raters being uniform in having no exposure.

Group Severity

Speakers were considered slightly more intelligible by experienced raters than by non-experienced raters. The difference is less than half a scale point, suggesting that experience does not have a large influence on intelligibility scores; nonetheless, the analysis demonstrates that experienced English language raters perceive non-native test taker speech to be easier to recognise than typical native English speaking listeners do. Such an outcome conforms to research suggesting that exposure to non-native speech improves processing speeds of such speech (Munro & Derwing, 1995b; Clarke & Garrett, 2004), and that rating experience results in leniency when judging criteria such as accentedness (Thompson, 1991), comprehensibility (Saito et al., 2016). It broadens research showing that when intelligibility is measured via a transcription task, experienced listeners find test taker utterances easier to recognise (Kennedy & Trofimovich, 2008; Bent & Bradlow, 2003; Munro et al., 2006), by demonstrating that the same applies when intelligibility is measured via a rating scale.

Why might experienced listeners find English speech that contains suprasegmental errors more intelligible than non-experienced listeners? Saito et al. (2016) argue that rater experience may lead to leniency due to experienced raters attending more to the message the speakers intend to convey, rather than surface-level characteristics that less experienced raters attend to (p151). This can be ruled out in the current study because intelligibility is primarily concerned with surface characteristics and understanding the message would have been incorporated into the comprehensibility score, rather than the intelligibility score.

It is probable that the difference in scores is related to experienced raters having greater familiarity with L2 varieties of English than their non-experienced counterparts. In other words, they have

more effectively acquired L2 varieties of English and are more efficient at processing them. Exemplar models of speech recognition present a possible explanation for this. Bybee (2001; 2010) argues that a listener's lexicon consists of base words and their variants, but that long-term memory also includes exemplars of words and utterances that have been encountered in previous interactions. These words and utterances are retained in a variety of contexts and voices. Speech recognition then relies on a listener matching the incoming signal to the exemplars held in memory. Competition on the basis of similarity is central to theories of spoken word recognition where units compete against multiple candidate words for recognition (Luce & McLennan, 2004: 594); in Multiple Trace Theory, the retention of traces of utterances from speech encounters facilitates this process.

The greater a listener's exposure to different varieties of English, the more traces of speech they retain in memory. As such, intelligibility is emergent in the sense that it emerges from repeated exposure to different varieties of pronunciations of words and utterances. It is also serendipitous in the sense that it depends on what types of non-native speech a listener has been exposed to. In other words, recognition of a certain variety of English depends on traces of that variety being stored in long-term memory, which in turn depends on frequency of exposure to that variety. In this research, the experienced raters had more exposure to foreign accents and learning foreign languages and therefore were more likely to have acquired a lexicon containing broader exemplars of non-native speech. This in turn results in more accurate matches when speech input is compared to memory. Therefore, if we accept Multiple Trace Theory, it stands to reason that experienced raters would have an advantage over non-experienced raters in terms of both acquisition of non-native speech and on-line processing. In the language testing context and in the method employed for this study the effect snowballs due to raters not being able to interact with the speaker, i.e. being unable to ask for clarification.

Finally, the extent to which experienced raters reflected on the rating process presents an interesting challenge to interpreting the results of this research. Raters are known to use monitoring strategies when rating (Sakyi, 2000; Cumming et al., 2002), but insight from the qualitative analysis reveals the extent to which experienced raters may be moderating their scores to match their perception of how a less linguistically experienced listener may respond. This is exemplified in the following excerpt:

ER_15: perhaps living in Spain and hearing this stuff . perhaps I'm being over generous in terms of how this guy would come across to a native speaker who's not an EFL teacher. you're always sort of thinking about this

This may go some way toward explaining the relatively small difference in severity between experienced and non-experienced raters. This is examined in more detail in the next chapter but, in short, it illustrates the highly complex decision making process that raters employed.

Summary

Experienced raters have several characteristics which potentially make them more sophisticated listeners than the people a test taker may encounter in a real-world context. These characteristics include the fact that professional raters, by virtue of their experience, have a broad range of linguistic experience, and are familiar with listening for the purpose of assessment.

One aim of this research was to examine whether the way non-standard suprasegmental features are interpreted in the test encounter is consistent with listening outside the test context. The finding that the differences between experience groups was rather small tends to conform to Nichols' statement that assessing communication is not an "esoteric skill requiring arduous

specialist training” (Nichols, 1988: 14, cited in Barnwell, 1989: 6), at least insofar as a 9-point global intelligibility scale is concerned. This extends the validity and reliability arguments put forth in reference to RQ1. In other words, it endows an intelligibility scale used to measure suprasegmentally errant speakers with cognitive validity, by demonstrating that the cognitive demands on professional listeners are similar to those of typical listeners. Cognitive validity is defined as how effectively a given assessment task elicits the cognitive processes that take place in authentic interactions (Field, 2011: 65). By contrasting experienced raters’ responses to a performance against those of non-experienced raters, it provides evidence supporting the notion that the process of recognising words and utterances that experienced raters undertake in a speaking test is similar to the way typical listeners do the same.

Chapter 5: Qualitative Findings and Discussion

The purpose of the qualitative analysis was to answer RQ3 relating to the terms raters attend to when making intelligibility judgements. The raw data comprised of 30 transcriptions of responses to a stimulated recall procedure. In this chapter the analytical approach is described in Section 5.1, and the findings are then reported in Section 5.2 and discussed in Section 5.3.

5.1 Qualitative Analysis

The audio recordings of the stimulated recall procedures were processed in a tape-transcribe-code-interpret cycle (Lapadat & Lindsay, 1999). The transcriptions of the verbal reports were combined with raters' written comments, as well as the researcher's field notes, and these were all coded and categorised.

5.1.1 Transcription

The audio recordings were transcribed using symbols from the Jefferson transcription system (2004):

- . a pause
- (()) contextual information and the researcher's field notes
- / rising pitch (pitch only marked when relevant to a particular point, e.g. on occasions when a rater was exemplifying a speaker's intonation)

\ falling pitch

(/wɜ:d/) phonetic transcription where relevant, symbols conform to The International Phonetic Alphabet on page 12 of this thesis.

underline emphasis

Some elements of speech were retained where useful: pauses were identified throughout; emphasis and intonation contour was signalled when required to understand the rater's point; and phonetic transcription was used in instances where raters imitated or exemplified. However, the researcher followed Lapadat's advice to "resist the temptation to create transcripts that will fit all needs" (2000: 215) and focused on producing transcriptions that clearly reflected what raters reported as motivating their scores.

An external auditor spot-checked the response to one test taker on each transcription. She questioned the convention used for transcribing pauses after checking the first transcription. Specifically, readability was being negatively affected by transcribing pauses too strictly. This was discussed and a new convention agreed for later transcriptions. Other minor points raised by the auditor were easily resolved. Following the transcriptions, credibility was supported using member checks. Three participants were asked to look over the transcriptions of their interviews and comment on whether they thought they were accurate. In all cases no issues were raised. Two transcriptions are reproduced in Appendix D, one from an experienced rater and one from a non-experienced rater.

5.1.2 Coding

The audio transcriptions were coded in NVivo 10 (QSR International, 2012). Initially a descriptive open coding strategy was employed (Saldaña, 2015: 61). Segments of the transcriptions were

assigned codes that made use of the raters' own language as far as possible. The disadvantage of an open coding strategy was that it was time consuming and resulted in a large number of codes, many of which were not relevant to the project, such as reference to grammar and vocabulary. However, it did allow the raters' own terms to dictate the coding, and permitted nuances of each rater's approach to feed through into the analysis.

It was not always possible to use the raters' own language to develop the codes. Certain linguistic features were communicated through imitation, exemplification, or by 'talking around' the feature of interest. These ambiguities, which are not uncommon in this type of analysis (Chi, 1997: 15), could only be resolved by researcher interpretation, for instance, in the following extract where NR_03 imitates the speaker's intonation.

NR_03: she's making very short sentences . noun or an adjective . the man has a /watch .
the lady has a /ring . she may be /married . ((imitating the speaker's rising
intonation))

In these cases, the researcher applied the code that was most appropriate, even though the rater did not refer directly to the feature of speech they were imitating.

The first round of open coding was undertaken twice. The researcher's reflective journal highlighted that, as the first round of coding progressed, the researcher became more experienced, the codes became more clearly defined, and the researcher's familiarity with the transcriptions improved. The process of constant comparison where newly coded segments of the transcriptions are compared to previously coded ones also resulted in a complex application of codes that changed through the coding procedure. Therefore, after coding the data once, it was deemed necessary to run through all the transcriptions again applying the same strategy.

Following the initial open coding, a round of focused coding was undertaken to organise the data into categories containing codes that were thematically or conceptually similar. Focused coding requires that the researcher identify the most salient codes and categories, to “determine the adequacy and conceptual strength of your initial codes” (Charmaz, 2014: 140). It requires the existing codes be refined and those codes irrelevant to the research problem to be discarded. This system of coding is associated with Grounded Theory, but is “appropriate for virtually all qualitative studies” (Saldaña, 2015: 213).

During the focused coding phase, it was possible to immediately discard certain codes which had emerged in open coding but which were clearly not related to pronunciation (e.g. ‘grammar’, ‘vocabulary’). Other codes were equally easy to retain due to their clear link to the research problem (e.g. ‘lilt’, ‘monotone’, ‘intonation’). These codes were reviewed to ensure they were applied consistently. There was found to be some duplication where different codes were used to cover the same phenomena (e.g. ‘musical’ and ‘sing-song’). The researcher resisted the urge to ‘tidy up’ such codes, not least because they could illustrate shades of meaning, but also because they allowed the raters’ voices to be heard throughout.

5.1.3 Categorising Codes

The codes were arranged into categories in NVivo 10. This process involved dragging and dropping a code to nestle it under a broader or otherwise related code. This is not primarily mechanical or technical but rather is “a process of inductive reasoning, thinking, and theorising” (Taylor et al., 2016: 168). It is iterative and, as such, the hierarchies were rearranged, merged and revised throughout the focused coding phase and well into the analysis. For example, ‘pausing’ remained a top-level code until relatively late in the focused coding phase. The

researcher read the text which had been coded as ‘pausing’ to identify precisely why it occurred to raters when making their judgements. This close reading showed that raters consistently linked errant pausing to issues of rhythm, as demonstrated in the extracts below:

ER_09: she pauses in just the right places . so she’s really got the rhythm of the language down to a T

NR_12: the pauses and the way that he delivers the words . even though they were very clear we tend to have a kind of rhythm to how we deliver a sentence . I just think it didn’t have any rhythm at all . it was just like what . is . the . next . word ((imitating the speaker pausing between words))

On the basis of this, ‘pausing’ was moved to the Rhythm category.

The outcome of the categorisation procedure was 64 codes organised into seven categories. These categories were organised into two groups as illustrated in Table 5.1 below. Group 1 includes suprasegmental categories and Group 2 includes codes associated with the suprasegmental aspect of test takers’ deliveries. These categories and the accompanying codes are listed in Appendix J, and they are discussed below in the Section 5.2.

Table 5.1: Coding Categories

Group 1	Group 2
Rhythm	Accent
Intonation	Infer Test Taker Characteristics
Lexical Stress	Irritation
	Natural

Group 1 categories were preconceived and devised primarily using a top-down approach (Chi, 1997: 25), i.e. based on the hypothesis that suprasegmental features of speech influence rater scores. They represent an inventory of the suprasegmental features of speech that influence rater judgements. Group 2 categories were not preconceived and emerged as relevant during the categorisation of codes. Group 2 categories are not clearly designated as suprasegmental but nonetheless, on interrogation of the data, they appear to be linked to suprasegmental features. For example, the researcher discovered that some raters encountered irritation caused by irregular rhythm and intonation, which in turn detracted from intelligibility. This is demonstrated in the following examples:

ER_12: it's more difficult for them to follow the rhythm of English and I suppose quite
 honestly I find it's slightly irritating

ER_05: it's not so much that it confuses just that it irritates . they talk about causing strain
 [...] and that's the kind of thing that makes it hard to follow the thread . you just
 think . why does he keep doing /this all the /time ((imitating rising pitch))

Coding quality was maintained by writing detailed analytical memos throughout all rounds of coding. The researcher also completed a reflective journal during the data gathering and analysis to identify any interpretational bias. Furthermore, an external auditor reviewed three of the coded transcriptions and discussed the codes with the researcher. The auditor made several comments on the codes which had been assigned but these were easily reconciled.

5.1.4 Analysis of Categories

Matrix queries were a primary tool in understanding how the data could answer the research questions. For instance, to understand how rater experience was related to rhythm, the rhythm codes were queried against rater experience and the output was a table showing the total number of times that experienced and non-experienced raters referred to each code. Relevant extracts of the transcriptions could then be grouped and viewed together, which facilitated comparison. Such an approach allowed recurrent themes and patterns to be identified in relation to test taker, rater, rater group, category, and code.

5.2 Qualitative Findings

The following description of the findings of the qualitative analysis is separated into four sections which relate the finding to: 1) the experienced raters; 2) the panel of phoneticians; 3) the non-experienced raters; and 4) rater familiarity.

5.2.1 Experienced Raters

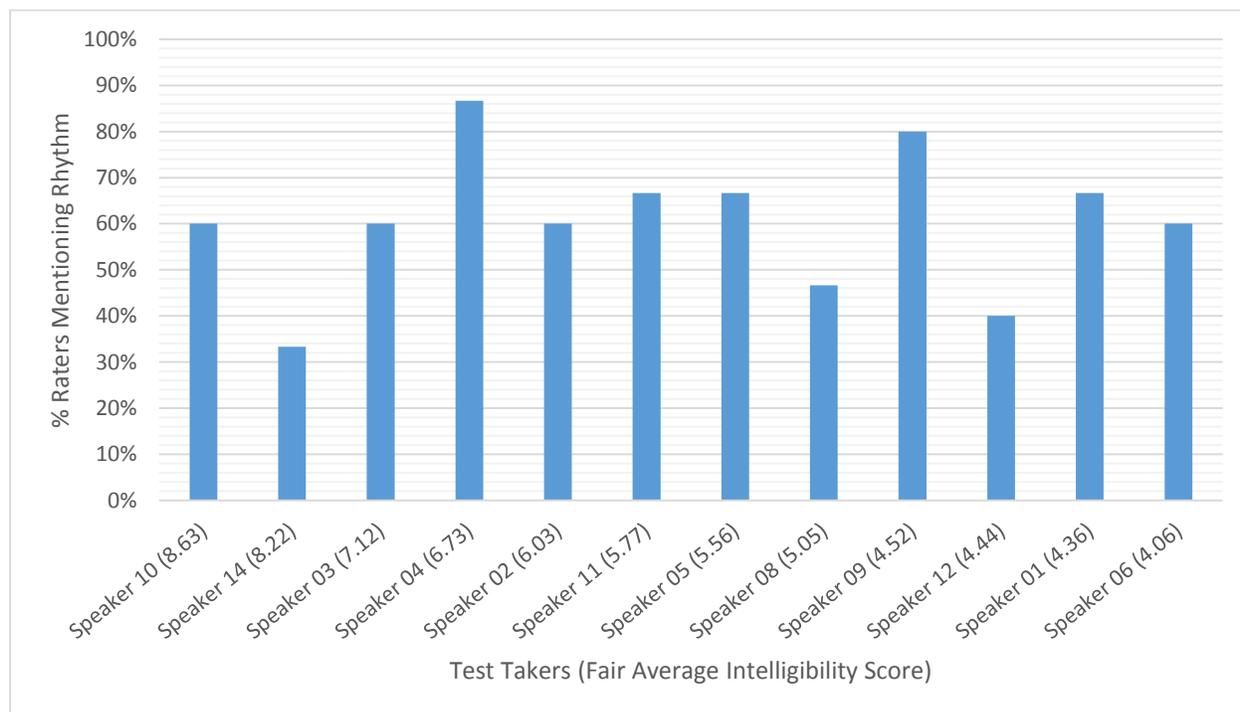
Rhythm was attended to most frequently by experienced raters, followed by intonation, and with lexical stress being raised least commonly. These are now taken in turn, after which the consistency of rater attention to these features is examined, followed by presentation of the specific suprasegmental terms raters used to refer to intonation, rhythm, and lexical stress.

Rhythm

Rhythm played an important part in the perception of intelligibility. It was commented on by more than half the experienced raters for all but three speakers (Speakers 8, 12, and 14). This is illustrated below in Figure 5.1, which shows the percentage of experienced raters who mentioned

rhythm for each speaker. The speakers are ordered from left to right descending by fair average intelligibility score as calculated in the quantitative analysis.

Figure 5.1: Percentage of Experienced Raters Mentioning Rhythm



Raters reported that for less intelligible speakers, rhythm could not emerge due to the quality and quantity of pausing and hesitation. The impact of this was delivery that was “jerky” (ER_06), “choppy” (ER_07) and “irregular throughout” (ER_04) for Speaker 9, or “jagged with uneven rhythms” (ER_04) and “staccato” (ER_06) for Speaker 6.

The way raters referred to rhythm for speakers they regarded as more intelligible was rather broader. For example, the rhythm of Speaker 3 was mentioned in the following terms:

ER_12: she’s French I think . but all the same the rhythm was fine

ER_01: I thought the rhythm was good . very good

ER_08: I thought it was really nice on rhythm

This highlights the primary way in which raters understood rhythm; they appeared to position a speaker's delivery on a spectrum from highly fragmented to rhythmical. Specifically, they did not discern different qualities of rhythmic speech, only the degree to which the delivery was rhythmical.

A particularly interesting aspect of the way raters talked about rhythm was the occurrence of strain. Several raters described lack of rhythm as causing excessive forms of effort and irritation. Table 5.2 below shows that rhythm was discussed by experienced raters in proximity to strain on 19 occasions. This is approximately 30% of the times strain was mentioned and contrasts with intonation, where it was only mentioned three times.

Table 5.2: Instances of Rhythm and Strain Coded Together

Strain Codes	Rhythm Category
Irritation	4
Annoying	1
Concentration	5
Effort	1
Frustration	4
Strain	4
Total	19

The two extracts below illustrate the way rhythm and strain related to each other:

ER_12: I thought it was a bit halting and staccato in a typical . I'm guessing Chinese . way
 . and then he doesn't complete the ends of words and that makes it hard to follow

[...] he's kind of telescoping his speech in a way . and then with the staccato it takes some concentration to follow

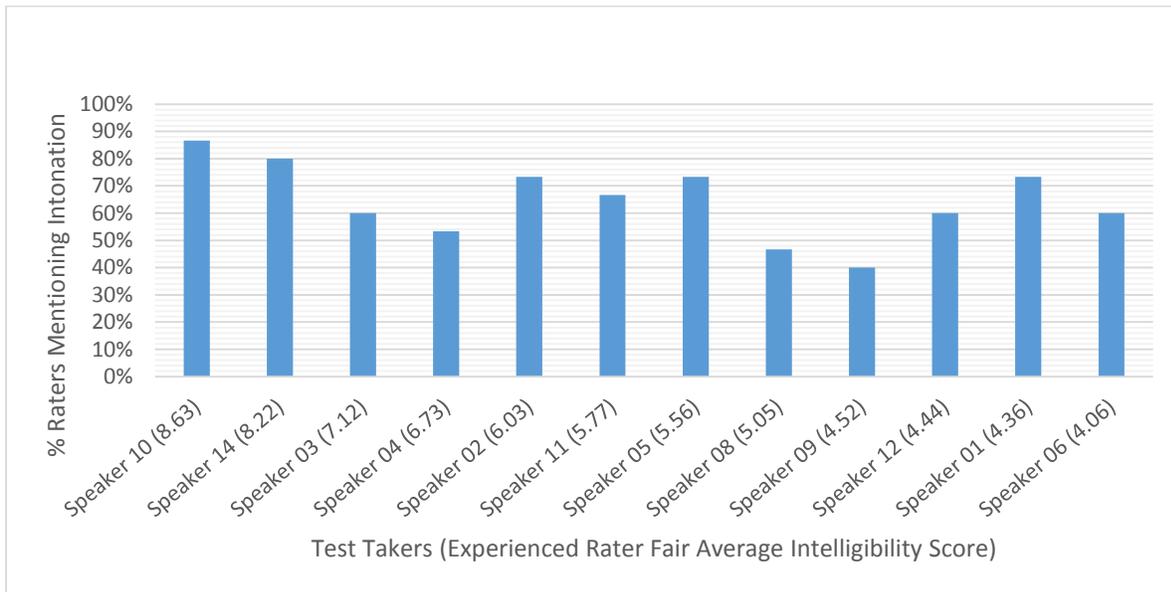
ER_02: it feels quite disjointed as well quite sort of fragmented . you're almost waiting for the next word . I don't know . it's like a young child reading . you know where they're putting their finger on each word so sometimes it's hard to get a global . you know . what is the sentence all about . because you're focusing on the individual sounds a lot more than you would normally . so I think it probably distracts you because your focus is almost in the wrong place as a listener

It appears that one mechanism in which lack of rhythm caused reduced intelligibility is the way fragmented delivery caused additional strain, which in turn resulted in difficulty recognising utterances.

Intonation

Intonation appears to be a common salient feature for the majority of speakers. Figure 5.2 below shows the percentage of experienced raters who mentioned intonation in relation to each speaker. The performances are ordered from left to right, descending by fair average intelligibility score. The chart shows that for all test takers, intonation influenced raters' judgements of intelligibility. For 10 of the test takers, intonation was such a pertinent feature of the performance that more than half the raters commented on it.

Figure 5.2: Percentage of Experienced Raters Mentioning Intonation (ordered by fair average scores)



There is a progression from poor control of intonation and monotonic delivery at lower levels of intelligibility to natural and native-like delivery at higher levels of perceived intelligibility. For instance, Speaker 1, who was regarded as the least intelligible speaker, was described as “fairly flat” (ER_02), “pretty flat” (ER_10) as well as being “off-key” (ER_07), and having “a problem with intonation” (ER_08). By contrast, good control of intonation was considered to have a positive impact on intelligibility. For example, Speaker 10, the most intelligible speaker, was regarded as having intonation that was “very nice” (ER_13) and “reasonable” (ER_08). ER_14 stated that the intonation “guides you through where you are in the sentence”.

One of the most striking features of intonation in the verbal reports was raters’ consistent use of nativeness and naturalness of intonation as a benchmark. Ten of the 15 experienced raters used the term ‘native’ in association with intonation in their stimulated recall responses and eleven used the term ‘natural’. For example, ER_02 describes how Speaker 1’s “unnatural intonation pattern for native speaker listeners” interferes with intelligibility. ER_01 stated that for Speaker 10 “the

inflection is almost like an English person”, ER_09 agreed that “she mirrors very well English language intonation so that helped”, and ER_11 regarded the speaker’s intonation as “almost native”. ER_03 makes a similar comparison in reference to Speaker 12:

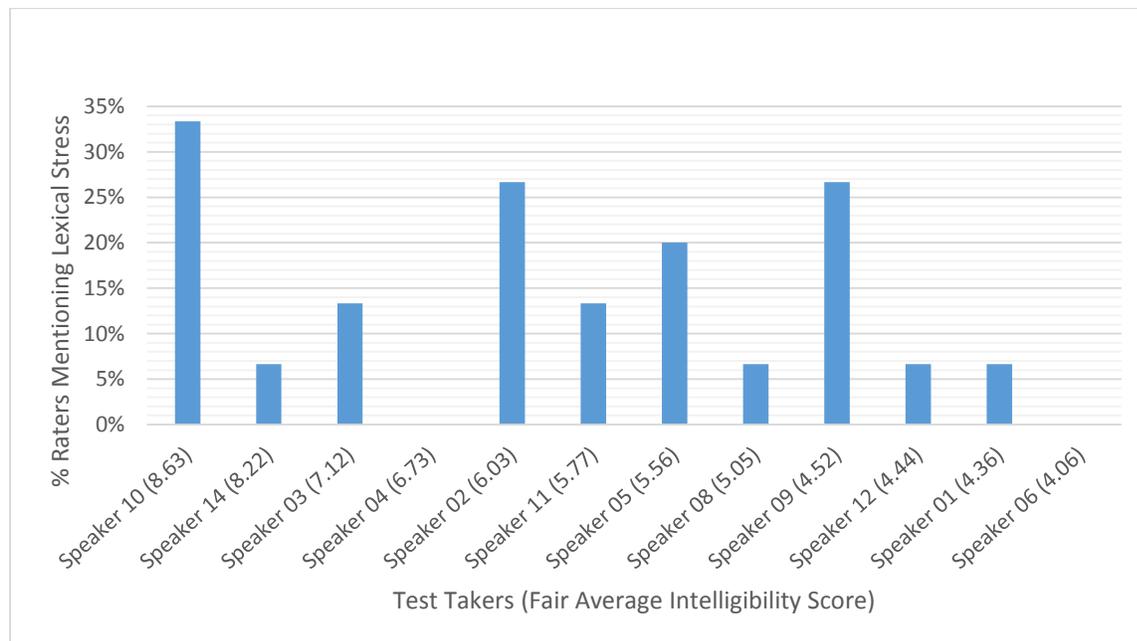
ER_03 a lot of the intonation was a little bit unnatural in terms of rising pitch . where I don’t think a native speaker would have that

Nativeness and naturalness therefore appear to be key ways in which raters judge intonation.

Lexical Stress

Lexical stress did not appear to be consistently influential. Figure 5.3 below shows the percentage of raters who mentioned lexical stress for each test taker. Lexical stress was never coded to Speakers 4 or 6, and it was never considered by more than 50% of raters to be influential to intelligibility for any particular speaker.

Figure 5.3: Percentage of Experienced Raters Mentioning Lexical Stress



Raters who were sensitive to lexical stress consistently identified it as an influential feature at the expected places (i.e. those outlined in the phonetic analysis: Speakers 2, 5, 8, and 9), but as Figure 5.3 illustrates, the proportion of raters who considered it influential to their decisions was relatively low.

Rater Consistency

Underlying the broad agreement on the roles of rhythm and intonation described above, there was some variation in the extent to which individual experienced raters were sensitive to suprasegmental features of speech. Table 5.3 below shows the number of times each experienced rater mentioned each suprasegmental category. ER_02 talked about suprasegmental features most extensively, with 48 references to these features. By contrast, ER_15 referred to them the fewest times, only referring to intonation and rhythm five and six times, respectively, and never mentioning lexical stress.

ER_05 and ER_15 attended to suprasegmental features the fewest times, but even so they referred to such features 17 and 11 times, respectively, indicating that they were clearly attentive to such features in the 12 audio recordings. Observation of their transcriptions indicate that they did not take a noticeably more holistic rating approach than other raters. Indeed, they discussed the performance in analytical detail in many instances. However, their stimulated recall interviews were shorter than other raters and they talked less about all phonological feature of the performance, indicating that they may have engaged with the task of reporting their memories less rigorously than other raters.

Table 5.3: Number of Times Categories Mentioned by Experienced Raters

Rater	Intonation	Rhythm	Lexical Stress	Total
ER_01	11	21	0	32
ER_02	22	25	1	48
ER_03	15	20	2	37
ER_04	18	18	6	42
ER_05	9	8	0	17
ER_06	15	18	0	33
ER_07	16	10	5	31
ER_08	15	12	5	32
ER_09	18	10	0	28
ER_10	18	9	3	30
ER_11	9	20	4	33
ER_12	10	11	1	22
ER_13	13	13	3	29
ER_14	18	7	0	25
ER_15	5	6	0	11
Total	212	208	30	450
Mean	14.13	13.87	2.00	30.00
StDev	4.56	5.94	2.17	9.13

Suprasegmental Terms

Experienced raters discussed rhythm, intonation, and lexical stress using a range of terms. Table 5.4 shows the codes categorised as rhythm. These are sub-categorised into three groups: 1) broad terms such as 'rhythm' and 'musical'; 2) reference to the breakdown in consistency of delivery with terms such as 'hesitation' and 'choppy'; and 3) technical terms such as 'weak forms'.

Table 5.4: Rhythm Codes used by Experienced Raters

Category	Rhythm		
Sub-category	1: Broad (29.8%)	2: Fragmentation (51.6%)	3: Technical (17.9%)
Codes	Rhythm (25.5%)	Hesitation (18.8%)	Linking (8.2%)
	Musical (4.3%)	Pausing (11.5%)	Weak forms (6.3%)
	Sing song (0%)	Staccato (7.7%)	Connected speech (3.4%)
		Halting (2.4%)	
		Stilted (2.4%)	
		Stopping starting (2%)	
		Choppy (1.9%)	
		Disjointed (1.9%)	
		Fragmented (1%)	
		Machine-gun delivery (0.5%)	
		Jerky (0.5%)	
		Segmented (0.5%)	
		Stalling (0.5%)	
		Broken(-up) (0%)	
		Jittery (0%)	
		Joined-up (0%)	
		Stutter (0%)	

0% denotes terms only used by non-experienced raters; values do not add up to 100% due to rounding

Primarily, raters talked about rhythm either broadly, e.g. ‘rhythm’, ‘musical’, which accounted for 30% of the times raters referred to rhythm, or in discussion of the characteristics of speech that stopped rhythmic delivery from emerging, e.g. ‘choppy’, ‘stopping starting’, which amounted to over half of all references to rhythm. Fragmentation as a source of rhythm breakdown was referenced by all experienced raters. Therefore, this represents the primary way in which raters attend to rhythm, or rather, a way in which they refer to the absence of rhythmic delivery.

The terms 'rhythm' or 'musical' were used by all raters and were assigned to all speakers at least once. This indicates that this broad approach to rhythm is consistently influential and is a major way in which raters interpret the concept of rhythm. Whereas 'musical' was used to talk about the delivery in positive terms, 'rhythm' was often used in the context of describing problems with rhythm. Either the speaker was perceived to be applying the wrong type of rhythm, as in the example below:

ER_04: she had quite a bit of hesitation and obviously very pronounced Italian rhythm transfer there

Or that there was no discernible rhythm, as in the following example:

ER_05: there is no rhythm as such . she's just saying people are going to the supermarket to buy things ((said in speaker's accent)) you know . but not even people are going to the supermarket ((said with exaggerated emphasis))

The technical terms in sub-category three, 'weak forms', 'connected speech', and 'linking', were applied moderately widely: 'weak forms' to eight test takers, 'connected speech' to five, and 'linking' to seven. But they were used by relatively few raters: 'weak forms' by four raters and 'connected speech' by three raters, although 'linking' was more widespread being used by nine raters. In addition, these terms were not applied consistently. There was never agreement by more than two raters that connected speech or weak forms were influencing the intelligibility of a particular speaker, and never more than three raters agreed that 'linking' was influencing a speaker's intelligibility. Therefore, unlike broad references to rhythm and reference to fragmentation, rater attention to technical fine details of the performance in relation to rhythm are not reliable indicators of intelligibility. Raters do not appear to attend to them consistently.

A final observation of the rhythm codes is the lack of references to rhythm timing. Raters were primarily concerned by the presence or absence of rhythmic delivery, not by the fine details of the type of timing. This is surprising, given the volume of research devoted to the role of rhythm timing to pronunciation proficiency.

Thirteen codes were categorised as ‘intonation’ and these were organised into three sub-categories as reported in Table 5.5 below. Sub-category one includes broad references to intonation and its synonyms, and accounted for the majority of times raters referred to intonation. Sub-category two includes descriptive terms to describe the delivery, and only non-experienced raters used these terms. And subcategory three is composed of technical terms describing elements of the speech signal that are associated with intonation.

Table 5.5: Intonation Codes Assigned by Experienced Raters

Category	Intonation		
Sub-category	1: Broad (61.8%)	2: Descriptive (0%)	3: Technical (38.2%)
Codes (per cent of all intonation codes)	Intonation (56.6%)	Bored (0%)	Focal Stress (26.4%)
	Inflection (2.8%)	Boring (0%)	Monotone (7.1%)
	Cadence (2.4%)	Bouncing (0%)	Pitch (1.9%)
	Tone of voice (0%)	Lilt (0%)	Tone (1.9%)
			Emphasis (0.9%)

0% denotes terms only used by non-experienced raters

The dominance of the terms ‘intonation’ and ‘focal stress’ is a primary finding of the qualitative analysis. It demonstrates that raters attended to intonation primarily in two ways: broadly in terms of acceptability; and narrowly in terms of suitability of stress. Pitch was a minor influential feature. Raters did attend to it in detail, as exemplified below, but not commonly.

ER_14: the . sort of intonation sounds quite natural . the sort of rising and falling or the pitch contour of the sentence sounds quite natural

Rater consistency in reference to pitch was reserved for reference to the delivery being flat or the pitch rising excessively. For instance:

ER_15 every time she's making a declaration she's going / up ((demonstrating))

ER_11: I think it's as if they are trying to confirm the meaning that they understand one particular word . he has a /ring and er er . the house was on /fire ((demonstrating speakers tone choice))

All 15 experienced raters used the term 'intonation', except for ER_01 who used the term 'inflection'. It is perhaps surprising that raters used the terms 'cadence' and 'inflection', given that they do not tend to appear in modern teaching or examination materials. Intonation can refer very specifically to pitch modulation, whereas 'cadence' might have connotations more closely associated with the global sense of delivery. Therefore, it is possible that the two raters who used 'cadence' and 'inflection' were trying to communicate a more global sense of delivery than the term 'intonation' implies. However, observing references to cadence tends to indicate they are being applied in a similar way to intonation, e.g.:

ER_09 he's got a good command of the language . but was a little bit monotonous and sometimes you had to get used to his cadence

ER_09 once I got used to that it was fine it was a bit choppy . but still it wasn't monotonous
 so I found listening to her was quite pleasant it had a nice rhythm going on . a nice
 cadence

Focal stress was used by all experienced raters, except ER_09, _14 and _15. When combined with 'emphasis' it accounts for 27 per cent of the references to intonation. Indicating that it is an important way raters make their judgements. Three examples of the way in which raters attended to focal stress follow:

ER_02: yea . she was emphasising some of the important words . so I got to the
supermarket ((with exaggerated stress on the first syllable)) . yea . sort of limited
 attempts at stressing words that were significant and again you can see that he's
 developing those skills

ER_04: there's a little bit of sort of staccato break there and . sometimes the sentence
 stress is uneven . which makes you sort of listen a little extra she has some of the
 extra syllables in . for example ice sculpture . it's ice sculpture /aɪsə skʌlptʃər/

ER_13: I think he's probably following the stress patterns the intonation the rhythm the flow
 of Russian . rather than . I found it difficult to understand . I think he needs to study
 how English flows . the rhythm the intonation etcetera . he's not using Russian .
 he shouldn't follow the Russian pattern

Monotonicity, pitch, and tone were less commonly mentioned by raters. There were only four references to pitch, and two to tone. This indicates that although some raters attended to these features, there was not enough consistency to make any clear generalisations. Monotonicity,

however, was raised by eight experienced raters, which suggests that most raters are aware of and sensitive to monotone delivery. Interestingly, monotonicity does not appear to be reserved for particular levels of proficiency. Although raters predominantly use it to refer to speakers who they consider to be relatively less intelligible, they also use it for speakers who they regard as relatively highly intelligible. For example, Speaker 14, who was the highest scoring speaker was described as “a little bit monotonous”, and “very monotonous” which is precisely the same language used to describe the lower scoring Speaker 6. It appears to be the case therefore that highly intelligible speech can still be monotonous at times.

5.2.2 The Panel of Phoneticians

The phoneticians were tasked with identifying notable errant suprasegmental features in each performance. The outcome of this provides a useful benchmark against which the raters' perception of the speech signal can be compared. It should be stressed though that the phoneticians do not provide a *correct* interpretation of the performances which raters might be expected to comply with. This is because phoneticians and raters have very different training, backgrounds, and ways of listening, and because the phoneticians had not been asked to judge the performances in reference to a particular pronunciation proficiency construct. Rather, the phoneticians provide an interpretation of the features of speech that notably deviated from L1 norms in each performance.

Raters were sensitive to all the suprasegmental features identified by the phoneticians except pitch range and vowel duration. Where inappropriate rhythm was identified as part of the phoneticians' analysis (Speakers 4, 5, 6, 9, 10, 11, 12) it was also described as problematic by raters. The instances where the phoneticians did not mention rhythm but the raters did are telling. For Speakers 1, 2, and 3, a minimum of 60% of the raters found rhythm to be influential to

intelligibility, but for those speakers the phoneticians did not mention rhythm. This underlines the importance of rhythmic delivery to perceptions of intelligibility even for speakers whose delivery is not considered, in phonetic terms, to be notably problematic rhythmically.

For intonation, there was a good relationship between rater judgements of monotonicity and the phoneticians' views on which speakers exhibited monotonicity. However, there was a more limited relationship for other characteristics of rhythm. This was because the phoneticians described the intonation of test taker performances primarily in terms of pitch movement, whereas raters discussed it more broadly in terms of nativeness and naturalness, with focal stress being the main detail of intonation that they attended to, and rising pitch being a minor feature that they noted.

5.2.3 Non-Experienced Raters

Non-experienced raters were less sensitive to suprasegmental features than experienced raters. Table 5.6 below shows that each suprasegmental category was mentioned fewer times by non-experienced raters than by experienced raters. It shows that experienced raters attended to intonation slightly more than rhythm, but that non-experienced raters attended to rhythm markedly more than intonation. It appears, therefore, that non-experienced raters were more sensitive to speech that is not delivered rhythmically than to speech which has poor intonation.

Table 5.6: Number of Times Categories Mentioned by Rater Group

Category	Experienced raters	Non-Experienced Raters
Intonation	212	117
Rhythm	208	155
Lexical stress	30	12

Although non-experienced raters mentioned rhythm less often than their experienced counterparts, they attended to it in a similarly broad way in positive terms, as in the below example:

NR_15: apart from a couple of minor words that she had trouble with that was excellent
 and it was easy to listen to [...] the nice rhythm throughout the whole thing

And in negative terms they also mirrored experienced raters in focusing on the impact of overuse of inappropriate pausing and hesitation, which resulted in rhythm not emerging:

NR_04: he's keeping up the speed [...] it's almost . there's almost a kind of a like a sort of
 staccato little breaks in that speed erm . it's sort of quite jittery

NR_11: I think if it's all over the place it's really hard to get . if it's disjointed then I think it's
 just . your brain . maybe it's just the way my brain works . I stop recognising things
 because I don't know where they're going with this

Intonation was approached in a way moderately consistent with the experienced raters. They discussed it in very broad terms as in the example below:

NR_03: intonation is more or less as it should be (.) there's nothing wrong with that

And they hold the same preference for measuring intonation against a native-speaker standard as experienced raters, such as in the example below:

NR_05: she was talking as we would normally talk with slight undulations of tone

However, because non-experienced raters were somewhat inattentive to intonation, there is less clear agreement as to its status. Table 5.7 shows the number of times suprasegmental features were mentioned by non-experienced raters for each test taker (descending by fair average). It shows that intonation was discussed in relation to all speakers, but that for certain speakers it was only rarely mentioned. For instance, for Speaker 4 it was only mentioned twice and for Speaker 5 five times. Rhythm received considerably more attention than intonation or lexical stress for most non-experienced raters.

Table 5.7: Number of Times Categories Mentioned by Non-Experienced Raters

Test Taker	Intonation	Rhythm	Lexical stress	Total
Speaker 10 (8.63)	9	9	0	18
Speaker 14 (8.22)	14	12	2	28
Speaker 03 (7.12)	7	8	2	17
Speaker 04 (6.73)	2	15	1	18
Speaker 02 (6.03)	15	15	2	32
Speaker 11 (5.77)	14	21	0	35
Speaker 05 (5.56)	5	16	1	22
Speaker 08 (5.05)	6	11	0	17
Speaker 09 (4.52)	7	16	4	27
Speaker 12 (4.44)	22	12	0	34
Speaker 01 (4.36)	7	7	0	14
Speaker 06 (4.06)	9	13	0	22
Total	117	155	12	284
Mean	9.75	12.92	1.00	23.67
StDev	5.51	3.96	1.28	7.30

Table 5.7 also shows that, in common with the experienced raters, lexical stress is clearly neglected. It was never mentioned in reference to half of the speakers, and the other half received few comments.

Although non-experienced raters were generally less sensitive to suprasegmental features than experienced raters, when they did reference specific features they were able to do so with a relatively high degree of complexity, e.g.:

NR_15: essentially it was monotone from my perspective . by being monotone there's no emphasis going on . there's nothing that would grip me in a conversation . nothing to make me really want to have a conversation with that individual . so it was that lack of emphasis . lack of change in pitch and tone that just made it bland

Non-Experienced Raters Consistency

There was moderate agreement on rhythm and intonation among non-experienced raters but there was some individual variation. This is illustrated in Table 5.9 below, which shows the number of times each rater mentioned each suprasegmental category. NR_11 was highly insensitive to suprasegmental features of speech, only mentioning them three times. By contrast NR_04 attended to them 38 times, more than many of the experienced raters.

This table shows that, with a few exceptions, non-experienced raters are well able to attend to suprasegmental features, indicating that being able to identify such features as influential to intelligibility is not simply the preserve of experienced raters.

Table 5.9: Number of Times Categories Assigned by Rater Group

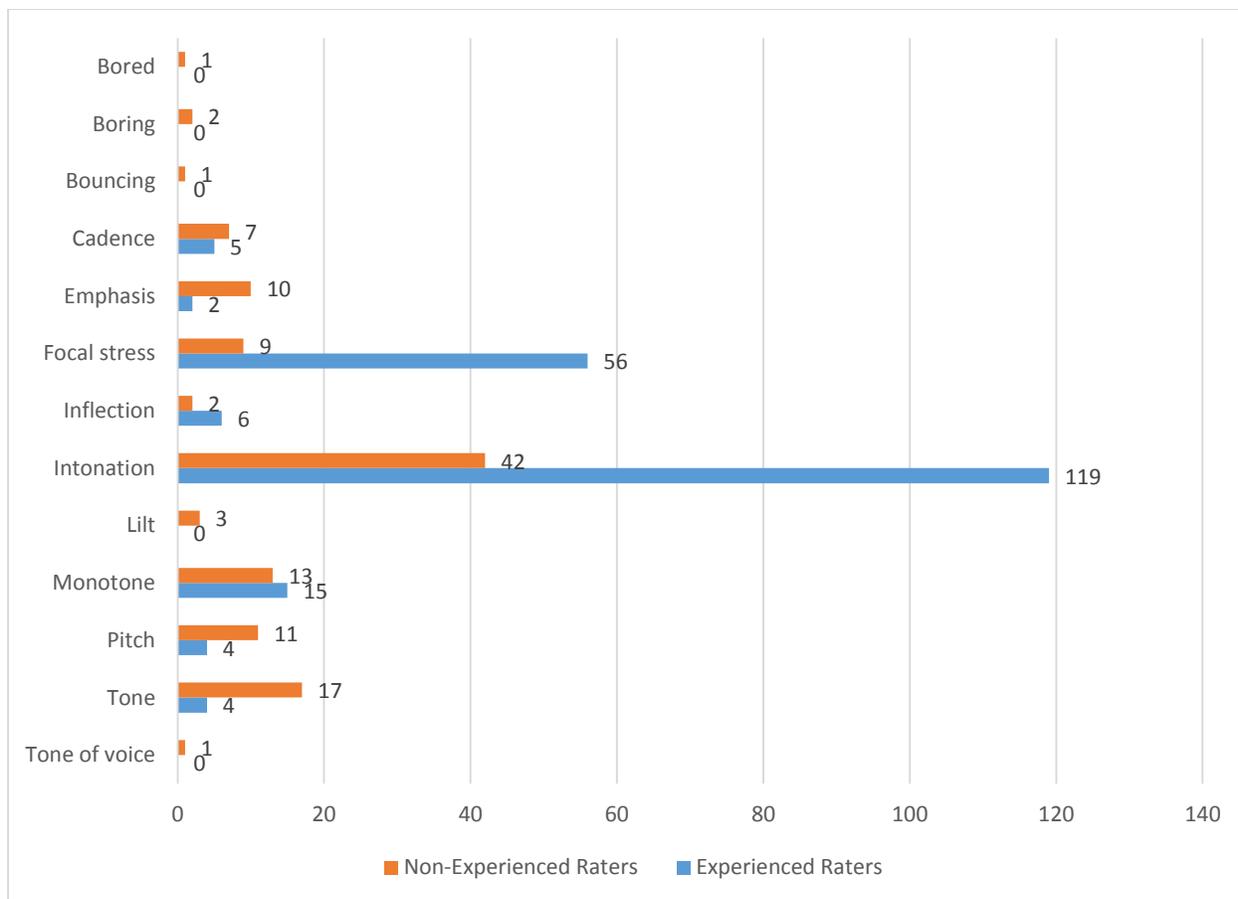
Rater	Intonation	Rhythm	Lexical stress	Total
NR_01	5	10	0	15
NR_02	9	15	1	25
NR_03	15	13	1	29
NR_04	13	23	2	38
NR_05	14	15	0	29
NR_06	1	7	0	8
NR_07	0	5	0	5
NR_08	4	6	1	11
NR_09	0	7	1	8
NR_10	12	7	2	21
NR_11	0	3	0	3
NR_12	3	6	0	9
NR_13	17	9	1	27
NR_14	9	13	3	25
NR_15	15	16	0	31
Total	117	155	12	284
Mean	7.80	10.33	0.80	18.93
StDev	6.26	5.38	0.94	11.08

Suprasegmental Terms

The different terminology used by experienced and non-experienced raters presents an interesting insight into the terms raters were able to employ to describe the performances. Counterintuitively the technical terms associated with intonation, ‘tone’ and ‘pitch’, were used more commonly by non-experienced raters than by experienced ones. For example, experienced raters mentioned tone four times (used by two raters) whereas non-experienced raters mentioned it 17 times (used by eight raters). Pitch was also mentioned by non-experienced raters more often than experienced raters. Figure 5.4 illustrates the number of times each code was mentioned.

Figure 5.4: Intonation Codes by Rater Group

Qualitative Findings and Discussion



The transcriptions suggest the non-experienced rater may be using ‘tone’ in a broader and less technical sense than the experienced raters. For instance, in the extract below an experienced rater makes a nuanced comment on the speaker’s use of tone:

ER_15: intonation . flat . there’s no real sense of variation in tone patterns which is again probably first language influence there

Whereas non-experienced raters tended to use the term to describe the speech much more holistically as exemplified below:

NR_10: it's a calming tone I guess . it's a more . it's a softer tone . it doesn't feel . you know when somebody's got a really loud brash voice and you feel like it's entering you're brain somehow in an invasive way . whereas hers didn't

This shows that even when non-experienced raters were apparently using rather technical terms, they were applying them less precisely than the experienced raters. In addition to this, non-experienced raters relied more heavily on descriptive terms, such as “lilt”, “bouncing”, and “boring” when talking about the intonation. This suggests that they were looking for ways to describe the suprasegmental aspects of the performance in the absence of effective technical vocabulary.

The implications of these differences are to demonstrate the value of rater training and experience. The experienced raters used more detailed and complex language when referring the performances. This indicates that they are likely to be able to interpret and apply more detailed descriptors with greater precision than typical listeners.

5.2.4 Rater Familiarity

A surprising discovery of the qualitative analysis was that experienced raters recognised and made allowance for their linguistic background. Thirteen experienced raters reported accounting for the fact that they had been in professional contact with a range of L2 speakers. They thought they would find a particular speaker easier to recognise than other listeners due to familiarity with the speaker's accent:

ER_01 another accent I'm quite used to I suppose . again so it's hard not to be a little bit biased

ER_02: for language teachers we're used to listening to the wrong intonation but again for native speakers it can be really really confusing . and so that sort of flat intonation . I don't know it could give the wrong message or it could make the message hard for native speakers to interpret I think

ER_10: he must be Russian and I'm quite used to listening to Russian people speaking so I'm quite familiar

The research was designed in such a way as to ensure the experienced raters had a broad range of linguistic experience, so it was not unexpected that raters encountered familiar accents. However, they also linked their familiarity with their perception of how an unfamiliar listener would respond to the speaker:

ER_15: perhaps living in Spain and hearing this stuff perhaps I'm being over generous in terms of how this guy would come across to a native speaker who's not an EFL teacher . you're always sort of thinking about this

ER_08: the word stress . the sounds weren't perfect but you could understand it without too much hassle . if you're used to listening to people learning English talking . I'm not so sure about other people

ER_02: I am very used to Italian so it's hard for me to be objective . but I think putting my mum's hat on she would struggle with that

This raises the possibility that in some cases experienced raters may not have been scoring according to their own sense of intelligibility, as instructed, but rather to some hypothetical sense

of intelligibility to the non-experienced listener. This may account for the rather small difference in scores between experienced and non-experienced raters.

Further support for this point is provided by the finding that none of the non-experienced raters mentioned accent familiarity, except two of the raters who were found to be borderline in terms of the Rasch IRT fit statistics when all raters were quantitatively analysed together. The selection of non-experienced raters was designed to minimise accent familiarity, so the fact that they reported being familiar with an accent may account for their misfitting scores. Interestingly, they make no suggestion of moderating their scores in light of this information, as the experienced raters had:

NR_15: she actually and I hope you don't mind me reporting this . she reminded me of somebody in my office who speaks in a very similar way

NR_04: the accent is familiar to me so that made it easier

It may be the case that the experienced raters were controlling their scores to bring them into line with what they expected to be a standard for a linguistically naïve native listener. This standard matched the decisions made by the majority of non-experienced raters, who were linguistically naïve, except for those raters who revealed familiarity with a particular speaker's accent.

5.2.5 Summary of Major Qualitative Findings

1. Raters were sensitive to a range of suprasegmental features of speech and were consistent with one another when attending to rhythm and intonation.

2. Rhythm was the most commonly referenced influence on intelligibility, followed by intonation. While several raters were sensitive to lexical stress and did identify it in the expected places, it was not mentioned consistently by raters.
 - i. Monotone intonation detracted from intelligibility and natural and native-like intonation promoted intelligible speech. Control of focal stress was influential to judgements of intelligibility but attention to pitch and tone was not common.
 - ii. Lack of rhythmic runs caused by hesitation and inappropriate pausing detracted from intelligibility, whereas speech regarded as musical or rhythmic promoted intelligibility. Rhythm timing and fine details of influences on rhythmic structure, such as weak forms, were not commonly attended to.
3. Non-experienced raters were broadly sensitive to the same features as experienced raters, although they attended to such features somewhat less commonly.
4. Experienced raters were aware of the impact of their language familiarity on their scores and might have been moderating their scores under the influence of this knowledge.

5.3 Qualitative Discussion

RQ3 concerned identifying the suprasegmental characteristics of test taker deliveries that raters attended to when making intelligibility judgements. The purpose was to establish the level of consistency among raters and thus the extent to which such features are suitable for inclusion in rating scale descriptors. For instance, if it had been found that raters did not attend to intonation consistently, then including intonation descriptors in a rating scale would be likely to promote inconsistent scoring.

RQ3 also involved investigating the features that non-experienced raters attended to when making judgements. The purpose of this was to establish how closely the interpretation of

experienced raters mirrored that of non-experienced raters, and therefore to examine the extent to which the resulting scale reflects typical non-specialist listeners. Finding that both experienced raters and naive listeners are sensitive to these features, even if the wording they used to describe them varies, indicates that certain suprasegmental features appear to be cognitively available to trained ears as well as to ordinary listeners. The outcome of this finding is that it is possible to develop mutually intelligible suprasegmental descriptors which may be suitable for scales, guidelines, and handbooks. The rest of this chapter is devoted to discussing the findings of the qualitative analysis.

A series of hypotheses were presented in Chapter 2 in relation to RQ3. It was anticipated that the following features would have an influence on rater judgements of intelligibility:

1. Intonation would be influential following existing research demonstrating that it influences rater decision making on certain criteria (Brown et al., 2005, Kang et al., 2010).
2. Focal stress was deemed to be less likely to influence intelligibility. It appears to not significantly influence processing of speech (Hahn, 2004).
3. The influence of lexical stress on intelligibility found using transcription tasks e.g. (Field, 2005), and laboratory experiments (e.g. Jesse et al., 2017) led to the hypothesis that this feature would follow through into the language assessment context and influence scalar judgements of intelligibility.
4. Rhythm was expected to influence rater judgements of intelligibility. There is a good link between rhythm metrics and proficiency judgements in the literature (e.g. Galaczi et al., 2012), as well as with syllable stress patterns and intelligibility as measured via a transcription task (Zielinski, 2008).

The outcome of the analysis in reference to these hypotheses will now be discussed.

Rhythm

The qualitative analysis indicates that the degree of rhythmicality in a test taker's delivery is a factor in speech intelligibility. The primary feature that raters attended to when making their judgements was rhythm. It seems that broken dysrhythmic speech presents a risk to intelligibility, as illustrated below:

ER_01 there's no sort of English rhythm to it is there . the rhythm is just very de de de
 ((imitating monotonic repetition)) . they don't stop talking they just keep talking and
 it's . you know whereas we would pause

ER_04: irregular throughout which makes it more difficult for listening to . I don't know if I
 listened again if I'd hear patterns . it was the irregularity which would make it more
 difficult

By contrast, the raters who took part in this study found speech delivered in rhythmic chunks to be rather more intelligible:

ER_02: she's breaking it up in quite a natural way isn't she . erm . it's not broken at
 individual word level or she's not struggling for words

ER_09 she pauses in just the right places so she's really got the rhythm of the language
 down to a tee

ER_08: despite the hesitations and the stresses she'd got the sentence rhythm

Recognition, therefore, appears to be more challenging when raters are not able to process chunks of speech but must instead process individual words.

Metrical structure might have an influence on recognition due to its role in helping a listener to identify word boundaries (Shoemaker & Rast, 2013; Cutler et al., 1997: 147; Cutler & Norris, 1988). Indeed, syllable stress patterns have been found to be influential to recognition (Zielinski, 2008). This is consistent with Galaczi et al. (2016), who found that low-proficiency speakers were more likely to have ‘unclear’ rhythm. Galaczi et al. (2012) found that 83% of raters surveyed regarded rhythm as salient to rater judgements. Similarly, Iwashita et al. (2008) found that low-proficiency speakers tended to have ‘unclear’ or ‘syllable-timed’ rhythm, and Tajima et al. (1997) found that native-like timing promoted intelligibility, which has clear parallels with the findings presented here, as raters often described native and natural delivery as promoting intelligibility. However, the raters who took part in this research were never concerned with the *wrong type* of rhythm, and timing was never discussed. Instead, rhythm was treated as a spectrum from present to absent. This is a notable finding, given recent research focusing on the influence of specific rhythmic features on scores (e.g. Kang, 2013a). Such research provides a great deal of insight into precisely which micro-features of rhythm influence oral proficiency, but this study shows that when placed within the language assessment context, raters do not consistently attend to the deliveries in such fine detail, at least at approximately levels B1 – C1 on the CEFR as represented in this study.

An alternative to the suggestion that rhythm influences intelligibility due to its role in delimiting words is the link to strain, irritation, and effort. The qualitative analysis demonstrated that fragmented speech caused strain and required effort, which in turn contributed to making it harder for the listener to recognise the words and utterances:

ER_02: yea it was kind of regular . there was an attempt to stress the main words . fairly fragmented . that would put a bit of strain

Irritation and effort as a reaction to non-native speech is a known phenomenon (e.g. Fayer & Krasinski, 1987) and it has been recognised in Band 9 of the pronunciation criterion on the IELTS Speaking Band Descriptors (see Appendix B), where speakers are characterised as “effortless to understand” (IELTS, 2012: 19). ‘Effort’ also appears on the CEFR phonological control grid (Council of Europe, 2001). Interestingly, Horner (2014), among others, (e.g. Ludwig, 1982), has suggested a suitable definition of intelligibility to be “degree of effort required of the hearer” (p115), noting that effort appeals to raters because it “reflects a perceived reality” (p116). Intuitively it makes sense: intelligibility is considered to be interactional; it requires the listener to pay attention, as well as the speaker to deliver well, and if the speaker is perceived to be delivering speech without sensitivity to the listener’s role, then the listener may find it harder to engage with what is being said, resulting in them recognising less of the speech signal.

The primary concern here is that irritation may be an attitudinal response to a speaker’s delivery, and may therefore prompt inconsistent responses from raters. In other words, it may be a function of characteristics of the listener beyond the ability of the speaker to articulate words and utterances in a recognisable manner. It is not wholly clear whether breakdown in intelligibility causes irritation or irritation causes the breakdown in intelligibility, so it may be possible that a rater’s sensitivity to strain interferes with their ability to recognise a test taker’s delivery to carrying degrees. More broadly, intelligibility is co-constructed, and as such listener attitude can be regarded as part of the intelligibility construct. That is, if a raters reliably disengages from a speaker due to their attitude about the performance, and if this is reflected in the general listening population, then it may be sensible to measure irritation as a legitimate source of breakdown of intelligibility.

The link found between fragmentation, effort, and reduced intelligibility seems to suggest that the mechanism by which limited rhythm disrupts perceptual recognition is that when a speaker's delivery does not fit into an expected pattern, this misfit breaks the listener's automaticity of processing the signal, thus causing effort and reduced intelligibility. This supports the supposition that a listener's brain is 'tuned' to receive a particular type of natural rhythm (Fujii & Wan, 2014). A telling statement from one of the experienced raters describes the implications of effective rhythmic delivery:

ER_13: it's not laboured at all . in the slightest it's not laboured . I could sit back just as if I was in a café with her . I didn't really have to stress . really focus on the words I could sit back relax and let her conversation flow over me . er like a natural conversation I guess . as you know if you're dealing with second language users often you have to really concentrate . you have to really focus . here no focus whatsoever . it was like listening to the radio

Therefore, regularity in delivery seems to fit the expectations of the listener, require limited or no additional effort, and allow the listener's well-practiced process of automatic listening to take place.

Several rhythm features regarded by raters to be influential to intelligibility in the current study have been reported as characteristic of fluency elsewhere. For instance, Brown et al. (2005) categorise 'hesitation' and 'repetition and repair' as subcategories of fluency, distinct from 'rhythm and stress', which are a subcategory of phonology, and in Rossiter's (2009: 403) fluency study, 'bad rhythm' is regarded as having an impact on the speakers' fluency. This suggests an interesting link between fluency and intelligibility. Fragmented speech limits fluency by definition

(McCarthy, 2010: 12), but it also appears to limit intelligibility. This was perhaps foreseen by Byrne (1986), who defined fluency as “the ability to express oneself intelligibly” (p9).

In Chapter 2, rhythm was characterised as a constellation of perceptual phonetic phenomena that may not be possible to measure objectively. A potential issue was raised that this conceptualisation of rhythm means it can only be measured subjectively and, as such, it is open to varying interpretations. This research has shown that despite the necessary subjectivity of measuring rhythm, it can be attended to reliably by raters.

Intonation

Speakers who were considered to be monotone and lacked control over focal stress, resulting in intonation that was not natural or native, were regarded as being relatively low intelligibility, e.g.:

ER_02: I think the major issue would probably be with the intonation pattern [...] a kind of unnatural intonation pattern for native speaker listeners because of the French

ER_12: a little bit halting . she stops after each phrase . not natural intonation

Speakers able to deliver speech with controlled intonation that was natural or native-like were regarded as relatively highly intelligible, e.g.:

ER_02: that naturalness again I think made it very easy to listen to . and then you can see . and then she was emphasising . you get a very /positive feeling and if you /win ((imitating the focal stress))

ER_05: native speaker like intonation

This focus on nativeness as a way of defining appropriate intelligibility presents a challenge to the interpretation of the results. It raises the issue of how nativeness is defined, and by extension the suitability of the term in the production of a suprasegmental rating scale. It is not clear what definition of native speaker raters were employing when they used the term. It is notable though that the terms 'native' and 'natural' often appeared together in the transcriptions, indicating that raters may have been using native speaker in a generic sense rather than specifically in reference to Received Pronunciation or any other specific variety of English. The primary purpose for retaining the terms 'native' in the scale derived from the data is that it is a term that raters themselves used consistently, and therefore appears to be a commonly understood term. An expansion of this study with non-native English speaking raters would provide better insight into the suitability of the term for the World English context.

The importance of monotonicity in breakdown of intelligibility is consistent with research by Pickering (2001), who employed Brazil's (1997) system of intonation analysis to argue that teaching assistants who spoke with flat and monotone pitch were not engaging to listeners. This is corroborated by Dang (2013), for instance, who found that the application of Vietnamese flat tone to English speech had a significant influence on intelligibility, as measured by a transcription task. The reason why monotonicity influences intelligibility may be rooted in the role of pitch variation facilitating speech processing. Indeed, Laures and Weismer (1999) argue that a reduction in intelligibility caused by presenting listeners with speech that had artificially flattened pitch profiles was due to pitch helping the listener parse continuous speech more easily. This is broadly consistent with the outcome of Cutler et al.'s (1997) review of relevant studies, although Cutler et al. retain the proviso that the relationship between suprasegmental elements of the signal and syntax is not one-to-one, nor is it completely conclusive (p170). The breakdown of

automaticity of processing may again be a factor in this reduced intelligibility. When recognition of the speech signal is automated, working memory can be dedicated to other processes (Field, 2003: 113). However, if the listener is required to engage conscious attention, for example when the speech signal has focal stress falling in unexpected places, then working memory must be engaged in recognising the delivery. If the delivery deviates critically from norms, the listener may not have enough capacity to recognise it all. The fact that raters refer to intelligible delivery as native-like is probably simply a way of saying that it allows them to process the speech with a high degree of automaticity.

Focal stress placement appears to be linked to intelligibility and this is consistent with studies addressing oral proficiency generally. For instance, Kang (2013b) found that a reduction in stressed words was accompanied by increases in proficiency. Such a link has not been observed in reference to perceived intelligibility before, although it makes intuitive sense. In the sample of test takers used in this study, in some cases stress is assigned to almost every word, resulting in persistent delivery of one-word intonation units. Speer and Blodgett (2006) attribute the below example to Pierrehumbert (without citation); it shows how misplaced stress can result in the intended words being completely mis-recognised, illustrating the way in which focal stress is linked to utterance recognition:

Triceratops oil Try, Sarah, topsoil Trice, air atop soil

Speer & Blodgett (2006: 506)

The example where every word is stressed ('try, Sarah, topsoil'), is different to the utterances with fewer focal stresses. Such coincidental ambiguity in meaning-linked stress is unlikely to take place in normal utterances, but nonetheless, excessive stress still requires the listener to apply effort in order to recognise the intended utterance. Focal stress has been found to not influence speech processing (Hahn, 2004) and judgements of attractiveness of speaker (Chen & Wang, 2016).

However, Hahn found that appropriate focal stress helped listeners remember information better and led to them judging the speaker more positively, a finding supported by Birch and Clifton (1995). Existing findings as to the role of focal stress in oral proficiency are somewhat tangled; this thesis presents evidence that this feature of speech is influential to rater judgements of intelligibility.

A surprise finding of this analysis is the way in which raters attended to specific pitch movements. Experienced raters were capable of analysing intonation in terms of specific pitch movements and tone choices. However, they did not do so commonly. The evidence of this study seems to indicate that some raters found pitch movement to be relevant, whereas others did not. It may be that consistency in responses to pitch movement was related to a characteristic of the listeners, such as the nature of their training, for example in phonetics, and even musical training, which is known to influence scores for accentedness ratings (Isaacs & Trofimovich, 2011).

A possible reason why rising pitch appears to be the most salient pitch movement in this study is the phenomenon of high-rising terminal, where a rising tone is used in statements in place of a more traditional falling tone. Six of the speakers were regarded by the phoneticians as employing excessive rising tones. Wells (2006: 37) argues that there may be a generational difference in how this tone is applied, suggesting rater demographics may influence its use and its acceptability to listeners. Wells goes on to caution English learners in use of high-rising terminal because it “may annoy older people” (p38), and indeed there were cases in the data of a rater demonstrating irritation at the excessive use of rising tone:

ER_05: it's not so much that it confuses just that it irritates . they talk about causing strain [...] and that's the kind of thing that makes it hard to follow the thread . you just think why does he keep doing /this all the /time ((imitating rising pitch))

ER_05: I don't find it much less annoying when native speakers do it . you know well I went to the meeting the other /day . and er they gave us three ballot /papers . to elect different member of the executive /council ((imitating rising tone on final word))

The age of the test takers is unknown, although Cambridge Assessment suggests that PET is suitable for students in late primary or secondary school, and FCE is suitable for students in secondary school (Cambridge Assessment, 2017c), indicating these test takers would be approximately 10 to 18 years old. The average age of the raters was 41. Therefore, it is possible that some of the speakers might have been employing a social accent which, while potentially suitable for their peer group, provoked irritation and limited intelligibility in the listeners who took part in this study. This has implications for the recruitment, training and standardisation of examiners.

The findings of this study support the trend indicating that intonation is influential to oral proficiency, and provide new insight into the assessment domain by demonstrating that these features influence raters' judgements of intelligibility. A range of studies have found that instrumentally measured suprasegmental features of speech are influential to intelligibility and other oral proficiency criteria. The findings presented here corroborate several of these studies while contrasting with others. In a study by Yates et al. (2011), intonation was the most commonly mentioned feature. Monotonicity in particular is a common way for raters to describe the speech signal appearing in relation to accent judgements (Hayes-Harb & Hacking, 2015) and fluency judgements (Rossiter, 2009). The important role of monotonicity found in Pickering's (2001) study is particularly noteworthy, since Pickering did not measure the impact on listeners but rather made her own interpretations of the impact. Equally, the fine level of detail to which Kang et al. (2010) correlated pitch movements to oral proficiency and comprehensibility found some parallels in this

study. They found high-rising and mid-rising pitch were good predictors of oral proficiency, and in the qualitative findings to this study raters did attend to pitch movement insofar as it was rising excessively.

The most notable deviation from the findings presented here are those of Derwing and Munro (Munro & Derwing, 1995a; Derwing & Munro, 1997). They found that their judgments of ‘nativeness of intonation’ and ‘prosodic goodness’ did correlate with rater judgments of comprehensibility and accentedness, but did not correlate to a transcription task designed to measure perceptual recognition. This may indicate that intelligibility appears to be less closely related to intonation and control over suprasegmental features in Derwing and Munro’s studies than they appear to be in this research. However, as was noted in Chapter 2, there is a known issue in getting reliable correlation statistics from data which has a lot of high values. In the case of Derwing and Munro, this relates to the high number of transcriptions that were scored as finding the speaker completely intelligible. It is probable therefore that Derwing and Munro underestimate the link between intelligibility and intonation.

Lexical Stress

Several raters were sensitive to lexical stress. They attended to it in the following terms:

ER_03: some words were not clear or the word stress was wrong on perfume . and married
not quite right

ER_04: the word and sentence patterns are fine

ER_07: there's very little incorrect sentence and word stress and I think generally the individual words are articulated clearly

ER_08: the internal stress of the words makes some of them more difficult to understand

The status of lexical stress in this study is puzzling. Clearly it has an impact on the ability to recognise words, as a range of studies have demonstrated. It has been characterised as able to “severely disrupt listener processing” (Richards, 2016: x), and intuitively it makes sense that placement of lexical stress should influence recognition because misplacement has a corresponding impact on segmental articulation. Yet only a few of the raters responded to it.

Furthermore, raters were rather less consistent in attending to lexical stress than other features. Lexical stress is known to result in reductions in intelligibility (Field, 2005) and is generally regarded to have a role in spoken word recognition, for example, in helping a listener search their mental lexicon for matches on the basis of stressed syllables (Dalton & Seidlhofer, 1994: 39). Thus, it is surprising to find that it was not more extensively mentioned by raters.

Possible reasons for the disparity between this study and other studies that measure the intelligibility of speech containing lexical stress errors, is the relative infrequency of lexical stress in the stimuli material. Specifically, the raters in this study listened to one minute of audio and in that minute might have encountered just one or two lexical stress errors. So, it may be the case that the relative importance of lexical stress was simply not high enough to overcome other persistent errors in terms of what raters perceived to be influencing their judgements. This is not a fault of the research method but rather a strength. The audio samples presented to test takers were authentic examinations and therefore the presence of such errors reflects the frequency of such features that a test taker may encounter in a standard testing context. In fact, the audio

samples probably over-represent such features since they were selected specifically to include the presence of suprasegmental errors. There were relatively few lexical stress errors in the speech samples, and thus the gravity of these errors was not high. This is compounded by the way different lexical stress errors have different error gravities (Field, 2005; Richards, 2016), meaning some of the errors delivered in the speech samples used in this study may not be damaging to intelligibility.

In addition to this, not all lexical errors detract from intelligibility in the same way. The direction in which lexical stress deviates from the norm may influence intelligibility (Field, 2005). Therefore, although the importance of lexical stress can be high, the relatively infrequent occurrence of influential lexical stress errors might have a low cumulative impact on intelligibility. Cutler asserts this when arguing that, although lexical stress can be used to establish syntactic category, the influence of broader context is likely to make the errors less critical in context (Cutler, 2015: 107-9; Cutler, 2005: 282).

Non-Experienced Raters

Non-experienced raters were broadly attentive to suprasegmental features in a similar way to experienced raters. They were somewhat less sensitive to such features and this is likely to be due to their lack of experience and training in language learning and testing contexts. The finding that they responded similarly to experienced raters in many ways indicates that assessing suprasegmental features of speech is not necessarily an esoteric skill, and does not need to be accompanied by extensive training. This provides empirical evidence supporting the inclusion of such features in scales. As was noted above, raters are less confident when it comes to assessing suprasegmental features; this research indicates that native English listeners have an innate and common understanding of the features that are influential and, as such, rating scale developers

can be less wary about including them in pronunciation scales. Furthermore, involving non-experienced raters provides inferences as to the pronunciation construct. Finding that non-experienced raters interpret the speech signal with a good degree of sophistication indicates that the existing pronunciation scales and frameworks may exhibit construct under-representation. Construct under-representation is regarded as a major threat to validity (Messick, 1989; Messick, 1992). In other words, this research demonstrates that suprasegmental features of speech are a component of the intelligibility construct in typical interactions, meaning appropriate coverage of the construct may require suprasegmental features to be a more prominent in pronunciation rating scales.

5.3.1 Summary of Qualitative Discussion

These findings relate to the aims of this research in five ways:

1. they provide detailed insight into precisely how listeners attend to suprasegmental features of speech;
2. they demonstrate that the suprasegmental characteristics of rhythm and intonation can be consistently attended to by experienced examiners;
3. they link experienced raters' judgements of such features to authentic communicative encounters due to the parallels between the way experienced and non-experienced raters attended to the speech signal;
4. they complement much of the research into the way speakers respond to suprasegmental features of speech by demonstrating how they are interpreted in reference to intelligibility in the language assessment context;
5. they provide sufficient agreement among raters to justify the development of a series of rating scale descriptors. A primary aim of this research was to establish whether it was possible to include suprasegmental features of speech in rating scale descriptors. Raters

were found to have a common perception of which features are influential to their judgements, and therefore that such features can be examined reliably by raters.

Existing intelligibility studies typically statistically associate instrumentally measured suprasegmental features with intelligibility, as it is measured using objective tasks such as transcription. Such procedures provide good validity arguments for ensuring raters take heed of suprasegmental features when making their judgements, but the methods do not emulate the language testing context particularly well. It is not clear how such technically measured features relate to perceptual features as defined by raters. As a result, operationalising the findings of such studies on a rating scale is difficult. As Lumley (2002) argues in the context of writing assessment, raters must come to some synthesis of the performance and the scale in order to achieve a score, it is not simply just a case of comparing the performance to the scale and expecting a consistent outcome.

The findings presented here extend the professional rater's perception to that of typical listeners. This indicates that the consistency exhibited by each experienced rater is not necessarily a function of their training or experience, but rather a common feature of native listeners. The finding that experienced raters were consistent with non-experienced raters endows the results with generalisability and cognitive validity in the sense that the relevant cognitive processes taking place among professional raters are consistent with those taking place among typical listeners. The fact that some raters discussed the features in more detail than others can be accounted for by the way in which some raters approach scoring holistically and others more analytically (Pollitt & Murray, 1996). Indeed, some raters have been found to approach scoring from a global perspective even when using an analytical rating scale (Joe, Harmes, & Hickerson, 2011).

In Chapter 2 it was suggested that the majority of current rating scales and frameworks are lacking in detail when it comes to suprasegmental features. For instance, although monotonicity was found to be highly influential to rater decisions according to the verbal reports collected for this research, they were not mentioned explicitly on any of the scales and frameworks reported in Appendix B. Equally, rhythm is often discussed in negative terms by raters who took part in this research, but rating scales typically only include positive descriptors. The impact of this is the potential for construct under-representation, which is one of the key threats to the validity of an assessment (Messick, 1989). In other words, the finding that these features are influential to rater and typical listener interpretations of intelligibility, but are not included in certain analytical rating scale descriptors, indicates that examinations using these descriptors may not be assessing the whole pronunciation construct. The impact of this is that such assessments might deliver scores which do not fully reflect test taker proficiency.

Chapter 6: Towards Suprasegmental Rating Scale Descriptors

The preceding chapters have described a situation in which raters attended to features of speech in a consistent fashion and scored speakers for characteristics of intonation and rhythm consistently. This indicates the viability of developing descriptors to measure suprasegmental features of speech, with the ultimate aim of making such descriptors available to designers of pronunciation rating scales to supplement the suprasegmental features of speech on existing rating scales and frameworks.

This chapter describes the approach undertaken to construct a sample set of descriptors representing suprasegmental features of speech. This involves defining the descriptor criteria and describing the way in which test takers were grouped to populate the scale points. Then the proposed descriptors are presented, justified, and discussed.

6.1 Approach to Developing Scale Descriptors

A rating scale consists of three fundamental elements: 1) the criteria, which typically appear at the top of a rating scale and define domains of proficiency; 2) the scale points, which typically run down the side of the scale and define a series of ability levels for each criterion; and 3) the descriptor text, which describes the ability for each criterion at each level.

For a series of scale descriptors based on the data collected in this study, the criteria were defined as the suprasegmental categories established during the qualitative data analysis: rhythm, intonation, and lexical stress. Rhythm and intonation were consistently treated as salient by raters who appeared to have a common understanding as to the influence of these features on intelligibility. Importantly, this common understanding was also held by non-experienced raters, which gives any resulting descriptor a link to authentic communicative encounters. Although lexical stress was found to be less influential to rater judgements than rhythm or intonation, several experienced raters showed themselves to be sensitive to errors in lexical stress placement. Lexical stress was mentioned 30 times in total by experienced raters (an average of twice per rater), whereas non-experienced raters only mentioned it 12 times in total (an average of 0.8 times per rater). Therefore, lexical stress appears to be rather more salient for the target audience of this study, raters and language teachers, and due to this it was included as a criterion in the suprasegmental descriptors. In addition to this, including lexical stress as a criterion may promote positive washback for the teaching and learning of lexical.

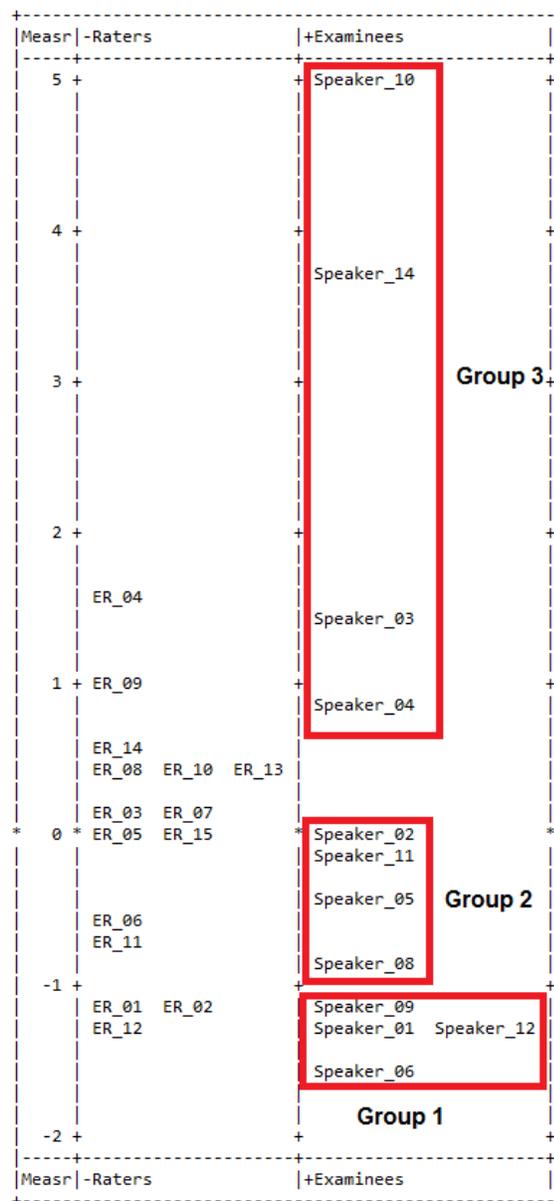
Producing the scale points of the descriptors required the performances to be grouped into common ability bands. This was achieved by using the fair average scores assigned by experienced examiners as part of the quantitative analysis of this study. This retained the study's focus on intelligibility and ensured the content of the scale was clearly linked to the scores assigned. The common ability bands were formed by creating three groups, each of an equal number of speakers. Table 6.1 shows that the average intelligibility score for each group progresses from 4.34 in Group 1, to 5.6 in Group 2, and 7.67 in Group 3. The score range also increases from Group 1 (0.46) to Group 2 (0.98) and Group 3 (1.9).

Table 6.1: Test Takers Grouped by Fair Average Score

Group	Test Taker	Experienced Raters Fair Average Intelligibility Score	Mean (Range) Fair Average Intelligibility Score
Group 1	Speaker 06	4.06	4.34 (0.46)
	Speaker 01	4.36	
	Speaker 12	4.44	
	Speaker 09	4.52	
Group 2	Speaker 08	5.05	5.60 (0.98)
	Speaker 05	5.56	
	Speaker 11	5.77	
	Speaker 02	6.03	
Group 3	Speaker 04	6.73	7.67 (1.9)
	Speaker 03	7.12	
	Speaker 14	8.22	
	Speaker 10	8.63	

The groups can be observed graphically below in Figure 6.1, which is the experienced rater intelligibility variable map output from the Facets analysis. It shows the estimated ability level of each candidate in logits. In this figure, it can be observed that Group 1 includes all the test takers who are under -1 logit, Group 2 includes test takers from approximately -1 to 0 logits, and Group 3 includes the test takers progressing from under 1 logit to 5 logits. The pattern of test takers being grouped at and below the average intelligibility level and then progressing up to five logits explains why Groups 1 and 2 are relatively close together in terms of their fair average scores and have similarly narrow ranges of scores. Group 3, however, has a somewhat higher average intelligibility score and a broader range of scores.

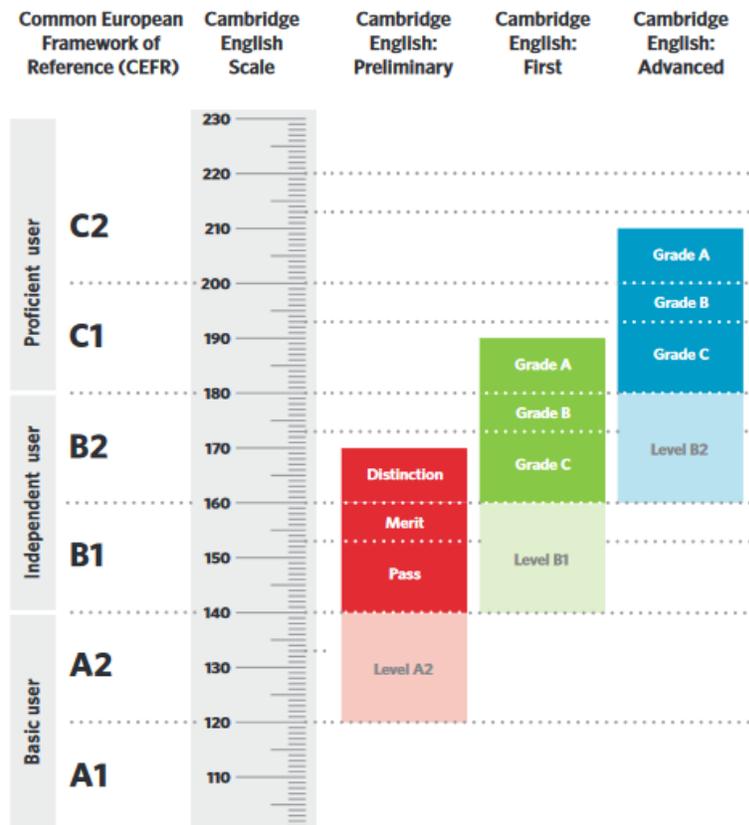
Figure 6.1: Experienced Rater Intelligibility Variable Map



In addition to using fair average scores, a preliminary link was made between the rating groups and the CEFR by referencing the original exam that the candidate had sat (PET, FCE, CAE). PET, FCE and CAE are related to CEFR bands B1, B2, and C1 respectively (North, 2005: 31; Taylor, 2011), as illustrated below in Figure 6.2. So, based on the original exam and pronunciation score that candidates received, it was possible to infer an approximate CEFR level for each group

of test takers. Cambridge Assessment provided the original pronunciation scores that had been assigned to the candidates by Cambridge Examiners. Scores for two candidates were unavailable but for the other candidates it was possible to demonstrate that none of the test takers had received a highest or lowest pronunciation score within the target band, indicating that they lay within the ability range targeted by the relevant examination. In other words, all the test takers with pronunciation scores were of an appropriate ability level for the examination that they sat.

Figure 6.2: Cambridge Examinations and the CEFR



(Cambridge English, 2015: 3)

Table 6.2 below shows the indicative link between test taker groups and the CEFR. It shows the possible range of CEFR levels the candidates might achieve, given the exam they took and the

score they received according to Cambridge English (Cambridge English, 2015: 3). An illustrative estimate of the CEFR level for each group is also presented in Table 6.2. The exam and Cambridge Assessment score columns indicate that fair average intelligibility scores are approximately mirrored by the progression from low-scoring PET candidates to high-scoring CAE candidates according to the scores for the pronunciation criterion assigned by Cambridge Assessment examiners. The relationship between fair average score and original pronunciation score is not perfect because the Cambridge pronunciation scores take account of more than just intelligibility. Furthermore, as Figure 6.2 above illustrates, there is expected to be crossover in ability level, meaning a high-scoring PET candidate can perform better than a low-scoring CAE candidate, for instance (Galaczi, French, Hubbard, & Green, 2011). This is purely indicative of the possible points around which each test taker might be performing. Nonetheless, it provides a useful framework in which to link the suprasegmental rating scale descriptors to the CEFR. These estimates would need to be confirmed by calibrating the descriptors using Rasch IRT. A proposal for such a calibration project is presented in Chapter 7.

Table 6.2: Scale Groups

Group	Test Taker	Exam	Experienced Raters Fair Average Intelligibility Score	Mean (Range) Fair Average Intelligibility Score	CEFR Range	Illustrative CEFR Level
Group 1	Speaker 06	PET	4.06	4.34 (0.46)	A2 – Low C1	B1
	Speaker 01	FCE	4.36			
	Speaker 12	PET	4.44			
	Speaker 09	FCE	4.52			
Group 2	Speaker 08	PET	5.05	5.60 (0.98)	A2 – Low C1	B2
	Speaker 05	FCE	5.56			
	Speaker 11	PET	5.77			
	Speaker 02	PET	6.03			
Group 3	Speaker 04	CAE	6.73	7.67 (1.9)	B1 – Low C2	C1
	Speaker 03	FCE	7.12			
	Speaker 14	CAE	8.22			
	Speaker 10	CAE	8.63			

The rating scale descriptors were constructed using the matrix query method of qualitative analysis described in Chapter 5. This method makes it possible to compare and contrast the codes and categories within each group of test takers to identify the most salient features at each level. Where patterns were identified, illustrative quotes were extracted from the transcriptions and placed on the relevant criteria and scale point of a visual display. This integrated display then exemplifies test taker performance at each level.

One more step was required to convert the visual display into a series of descriptors. A scale made up entirely of raters' own terms is appealing but can be unwieldy in a professional assessment context. This is due to length and variation in terminology, for instance. Fulcher's (1996; 1993) development of a fluency scale exemplifies this by resulting in descriptors containing up to 200 words, which are unlikely to be usable in a real-time assessment context. Therefore, the comments on each level of the visual display were summarised and abridged in such a way as to retain the meaning of what the raters said, as well as the actual words they used to say it, as much as possible. The outcome is a series of descriptors which have a suitable format for use in a live exam while retaining the authenticity of raters' own ways of understanding test taker speech.

6.2 Suprasegmental Rating Scale Descriptors

The integrated display that was developed by extracting representative quotes at each level of the scale is presented in Appendix K. Initially there were three scale points, one referencing each examination at levels B1 to C1 on the CEFR. However, the data was rich enough to discern a higher level in each category, ultimately resulting in six bands. Table 6.3 shows the final descriptors which were devised by transforming the raters' comments that populated the integrated display into an appropriate rating scale format.

Table 6.3: Suprasegmental Rating Scale Descriptors

Category	Rhythm	Intonation	Lexical Stress
B1	Highly fragmented delivery which sounds staccato or stilted and causes the listener to strain.	Flat intonation resulting in speaker sounding monotonous. Emphasis falls in unexpected places	Lexical stress is incorrectly placed on some words
B1+	Delivery is still somewhat fragmented. The speaker sounds staccato or jittery, and still requires concentration on the part of the listener.	As B1	As B1
B2	As B1+	Excessive unnatural rising intonation may make the speaker sound like they are listing, or cause a 'sing-song' delivery. Limited ability to stress important words.	Lexical stress is occasionally inappropriate and causes some words to be difficult to recognise
B2+	Effective control over rhythm beginning to emerge but in a limited manner. Hesitation is persistent causing the speaker to sound choppy.	Control over weak forms and application of some appropriate focal stress results in the speaker beginning to sound as though they are using effective intonation.	As B2
C1	Limited periods of fragmented delivery make the speaker sound choppy at times but generally the speaker breaks up the delivery in a natural way.	L1 interference persists although the speaker is now beginning to deliver periods of natural intonation. Focal stress is effective at times.	Word stress is generally appropriate and accurately placed on the whole.
C1+	Appropriate pausing and limited hesitation result in natural sounding rhythm most of the time	Speaker's intonation is natural and native-like or nearly native-like.	Very little incorrect word stress.

The descriptors demonstrate an intuitively sensible progression from broken, monotonous delivery at the lower levels to natural and native-like delivery at the higher levels. Rhythm is highly fragmented until B2+, where speakers begin to apply it effectively despite hesitation at times. There is a fine distinction between B1 and B1+, where listener strain changes to excessive concentration in the face of the delivery not being completely fragmented at all times. Hesitation persists after B2 all the way up to high C1 speech begins to sound native at B2+, then generally uses natural rhythm despite some fragmentation at C1, and finally sounds natural most of the time at C1+.

Progression of intonation ability is similarly sensible with monotonicity and lack of control of focal stress characteristic of B1. Then at B2 speakers strengthen their control over focal stress. This is particularly bolstered at B2+ where control of weak forms emerges. Additionally, it appears that chunking begins to take place more effectively at this level, resulting in the listener perceiving more complete intonation contours. This is related to rhythm, which also appears to begin to emerge effectively at B2+. Monotonicity may become less marked and by overuse of certain pitch patterns at B1. There is no mention of this phenomenon at B2+, although it emerges again at C1 in terms of L1 interference. In spite of this, the speakers at C1 are beginning to deliver periods of natural intonation, and by C1+ the delivery of intonation is native or near-native.

The lexical stress criterion is represented in less detail than the other criteria. As discussed in the qualitative findings, raters attended to lexical stress less commonly than intonation and rhythm. At B1 it may be placed incorrectly on some words, although by B2 the errors may be less common and have a less fundamental influence on intelligibility. By C1 lexical stress is generally appropriate and accurate, although even at C1+ occasional errors may persist.

6.2.1 Descriptor Development

This section describes how the integrated display was used to formulate the suprasegmental descriptors.

Rhythm

For the Group 1 test takers, raters regarded their delivery as hesitant and beset by unnatural pausing; they stated that it was “very stilted”, “very hesitant”, and felt that “pauses interrupted the natural flow”. Raters also made regular reference to a staccato effect, or to the delivery being stilted. Following this, the B1 descriptor employed the terms ‘staccato’ and ‘stilted’. ‘Fragmented’ was also used as a term to describe performance at this level. Several raters used this term, indicating that it is a recognisable phrase to use, but it also seems to capture the range of terms raters used to talk about how the delivery lacks any kind of rhythmical structure. For Group 1 speakers, certain parts of the performances in relation to rhythm are considered less damaging to intelligibility, and these feed in to the B1+ descriptor. At this level raters were still concerned about staccato delivery, referring to “staccato, little breaks” and regarding it as “quite jittery”. Equally at B1+ raters described concentration in reference to rhythm: “it required concentration [...] a little bit broken”, rather than the strain exhibited at the lower level (e.g. “there was a bit of strain on me”). Also at this point, raters more commonly referred to hesitation as the cause of this broken delivery, e.g. “I think the hesitation makes it quite difficult to follow [...] it’s choppy”.

There was not enough new discussion from raters at level B2 to create a separate descriptor. Raters in Group 2 were still concerned primarily about “fragmentation” and delivery that was “disrupted by pauses”, where rhythm does not emerge due to “quite a bit of hesitancy”. At the higher ability level in this group, however, it was possible to discern that raters considered test takers to be beginning to control hesitation. For instance, raters noted at this level: “it’s a bit

choppy but I still think she's obviously used to native English speech patterns because she can follow them", and "a little bit of hesitancy but she always managed to fill it in with an appropriate connective". This was operationalised on the descriptor as "Effective control over rhythm beginning to emerge but in a limited manner", which encompasses the break away from persistent fragmentation to a delivery which has notable periods of consistency.

For Group 3, the raters began talking about rhythm in positive terms despite still having some periods of limited rhythmic delivery. For example, the rhythm is described as "nice rhythm going on, nice cadence" and the raters regard it as natural. The descriptor therefore states that fragmentation, specifically 'choppy delivery', persists but there are now periods of natural delivery. The magnitude of positive terms increases at the higher end of C1. Raters argue that the speakers are "hardly hesitating" and have "got the rhythm down to a tee". Pausing is mentioned earlier in the scale and it emerges again at C1+ where raters regarded the pausing as appropriate. They also state that a speaker is "hardly hesitating", leading to a descriptor at this level referring to appropriate pausing and limited hesitation.

Intonation

For speakers in Group 1 raters complained that the delivery was "quite flat" and "monotonous", and described the emphasis as "being on words you wouldn't expect emphasis to be on". This fed into the descriptor at B1, which describes the language at this level as: "Flat intonation resulting in speaker sounding monotonous. Emphasis falls in unexpected places". It is difficult to discern a different level of performance at B1+ where raters continue to state that the performances were "flat" and "all monotone".

In Group 2, raters still stated that speakers were not using intonation to create focus: “there’s no attempt at emphasising really important words, so it does make it difficult”. Raters also begin to find that speakers applied tones but these are not well controlled or targeted. They complained about “unnatural rising intonation”, for instance. This led to the term ‘unnatural’ being used in the descriptor, as well as the reference to the delivery sounding like the speaker is repeating a list. At the higher end of B2, speakers begin to gain control over intonation. Raters stated that “there was an attempt to use intonation”, “he was using weak forms very naturally”, and “she was emphasising some of the important words”. These terms are therefore employed in the descriptor at B2+ which states: “Control over weak forms and application of some appropriate focal stress results in the speaker beginning to sound as though they are using effective intonation.”

In Group 3, the descriptor was constructed in reference to raters’ attention to delivery being “very natural”, and the focal stress being helpful: “in some places the sentence stress helped us, but he didn’t have a very wide range”. However, there is some discussion of the speakers’ L1 accents interfering with the intonation: “she had this you know sort of French intonation going up, etcetera”. At the higher level of C1, the descriptor describes performance as natural and native. This reflects the raters’ reference to “the inflection is almost like an English person” and “very nice intonation”, and “natural cadence”. The outcome is the descriptor: “speaker’s intonation is natural and native-like or nearly native-like.”

Lexical Stress

At Group 1, lexical stress is regarded as incorrect on some words. Raters stated that “the word stress was wrong on perfume” and “word stress is not particularly good”. This is reflected by a short descriptor at this level: “Lexical stress is incorrectly placed on some words”. It was not possible to separate the B1 group into a higher and lower set of test takers performances, but the

issue persisted at level B2, with the rater's statement that lexical stress is "occasionally inappropriate" and "the internal stress of the words makes some of them more difficult to recognise". The descriptor at B2 was therefore "lexical stress is occasionally inappropriate and causes some words to be difficult to recognise". At C1, raters used terms such as "generally appropriate" and "accurately placed". These are directly transposed into the descriptor: "word stress is generally appropriate and accurately placed on the whole." At the higher level of C1 it is possible to discern speakers who make very few or no lexical stress errors, and once again the descriptor is made up almost completely by the raters' terms, "there's very little incorrect word stress".

6.2.2 Descriptor Quality

The reader will have noticed that these descriptors deviate to some degree from descriptors typically found in language assessment scales and frameworks, primarily in reference to the use of negative descriptors. The Common European Framework of Reference for languages makes five suggestions for developing quality descriptors (Council of Europe, 2001: 205-7). They should be:

1. Positive
2. Definite (not vague)
3. Clear
4. Brief
5. Independent

The descriptors are certainly independent. However, they are not always positive and issues remain around the question of vagueness, clarity, and brevity. These quality criteria are now discussed in reference to the proposed suprasegmental descriptors.

Positive

Descriptors are often couched in positive language. The purpose is to focus learners and assessors on achievement rather than deficiencies. This is typical of user-oriented scales (Alderson, 1991) because it aligns them with learning objectives and promotes learner motivation, although Takala (2010), for instance, has argued against this, stating that including a speaker's constraints on a scale is regarded as useful by teachers. Several scales do not use exclusively positive descriptors, such as the TOEFL iBT integrated and independent speaking scales (ETS, 2014; Appendix B) and the PTE-A scales (Pearson, 2017a: 24; Appendix B).

The purpose of the descriptors presented here is not motivational or for language learning, but rather for valid assessment. The descriptors were derived from assessor language and thus they constitute assessor-oriented descriptors. Furthermore, the evidence gathered from the qualitative element of this research study overwhelmingly indicates that the way suprasegmental features of speech influence raters' perceptions of intelligibility is in negative as well as positive terms. Raters characterise low-level suprasegmental proficiency as broken, hesitant delivery that is monotonous and has irregular and oddly placed focal stress.

Definite and Clear

All rating scales have a degree of subjectivity, but the suprasegmental descriptors presented here provide a clear idea of the role of suprasegmental features at each proficiency level. Nonetheless, there are three instances of descriptors relying on modifiers that may result in problems of rater standardisation. These tend to relate to lexical stress, with terms such as "generally appropriate", and "very little". It must be noted that these are precisely the types of terms raters use when describing the language. In the whole qualitative dataset, use of 'very' is common, and the term 'quite' appears 560 times, so these modifiers are clearly important to the way raters characterise

the impact of suprasegmental features on intelligibility. A concern is the very fine difference between lexical stress descriptors at B1 and B2. Table 6.4 shows the way lexical stress was described by raters insofar as it was reported on the integrated display. It also shows the resulting suprasegmental descriptors which were derived from the integrated display. Use of the word “occasional” at the B2 level suggests raters regarded the impact of lexical stress to be somewhat less severe at this level than at B1. The difference is very nuanced, though, possibly meaning that raters may not be capable of operationalising it in a live test.

Table 6.4: Lexical Stress Descriptors

Level	Integrated Display	Suprasegmental Descriptor
B1	<p>“some words were not clear... the word stress was wrong on perfume“</p> <p>“word stress is not particularly good”</p>	Lexical stress is incorrectly placed on some words
B2	<p>"the internal stress of the words makes some of them more difficult to recognise"</p> <p>"occasional inappropriate stress"</p>	Lexical stress is occasionally inappropriate and causes some words to be difficult to recognise

The level of detail and the clarity of descriptors has been found to cause problems for raters (e.g. Yates et al., 2011), but the descriptors presented here are generally clear and precise. The terms used, such as ‘fragmented’, ‘monotone’, ‘intonation’, and ‘native-like,’ are all words which were reliably applied by raters to the performances, suggesting that they would be able to apply them when examining. If any words on the scale can be regarded as jargon, then such a term is common enough to the raters who took part in the study to be applicable in the target context. Resolving the issue of whether raters respond negatively to the terms used in these descriptors requires further validation. Usability testing, such as that exemplified by Harding (2016), would be a suitable way to identify precisely how clear raters find these descriptors to be. A proposal for such a procedure is presented in the conclusion of this thesis in Chapter 7.

Brief and Independent

The descriptors are independent and are relatively short – the longest is 32 words. The motivation for this research was to provide more detail on precisely which features are influential, so the additional length of the descriptors is inevitable. In fact, the development of the descriptors presented here already results in some loss of nuance, as can be observed by the difference between the extracts presented in the visual display in Appendix K and the resulting scale. The CEFR recommends that a descriptor becomes unwieldy once they are longer than a two-clause sentence and notes that, during the development of the CEFR, teachers tended to prefer descriptors shorter than 25 words (Council of Europe, 2001: 207). Some of the descriptors presented here are longer than 25 words and contain more than two clauses. As with understanding precisely how raters interpret the lexical stress criterion, usability testing is required to identify whether raters are able to operationalise them in a testing context.

In summary, with the caveats described above, the descriptors presented here conform adequately to the standards of quality outlined by the CEFR for all criteria except the requirement for them to be positive. The goal of this research is to develop a series of descriptors which can complement, or even supplant, the descriptors currently used by pronunciation scales and frameworks by investigating the way raters interpret performances. Raters interpreted the performances in negative terms, and so this must follow through to the final analysis.

6.3 Discussion

The suprasegmental rating scale presented in this chapter responds to Horner's (2013) suggestion that there is insufficient evidence on the role of intonation to place it on a rating scale. It also takes steps towards fully describing the role of rhythm, intonation and lexical stress as a speaker progresses through B1, B2, and C1 CEFR levels. The resulting scale represents features

of speech that have been shown to be clearly perceptible and salient to raters when making judgements. The resulting descriptors have several notable characteristics: they include certain features and terms that are not included in existing scales and frameworks; they are generally recognisable by expert raters, who were capable of expressing them in a way that is shared by other raters; and they relate to the perception of typical listeners. They include terms that may not be commonly used by applied linguists, such as ‘choppy’, but these are terms that raters used themselves when describing the language, and as such, they endow the descriptors with ecological validity. In other words, the descriptors would not simply impose a way of judging on examiners who use them, but rather they reflect the community of judges’ common understanding of how such features influence intelligibility.

A primary strength of these descriptors is that they describe features identified as relevant to both non-experienced raters and professional raters, even though the language used by non-experienced raters was somewhat less precise. This is not to suggest that naïve listeners would be able to use the scale as effectively as experienced raters, but it does provide external validation in the sense that a test taker’s developing mastery of rhythm and intonation is not something that can only be observed by experienced listeners, but is evident to untrained and inexperienced listeners as well.

A potentially controversial element of the descriptors is the inclusion of native and near-native levels of control over intonation and rhythm. Such terms are a common way for raters to characterise good control over intonation (e.g. Brown et al., 2005; Hayes-Harb & Hacking, 2015). Nonetheless, their inclusion in the descriptors may raise concern because it references native speakers as exemplars for proficient English speech. For instance, Jenkins (2014) argues that native-speaker norms have a reduced role in typical English-language interactions, and Davies (2011) criticises the notion of *the native speaker* as an idealisation (p306). Reference to

nativeness does not appear on any of the existing scales reviewed in Chapter 2, and presented in Appendix B, and most rating scales and frameworks avoid reference to native speaker competence (Taylor, 2006). The reasons for this may be a focus on cultivating broad intelligibility over accent reduction in pronunciation teaching (Levis, 2005). Several researchers argue that using nativeness to judge accent is wholly inappropriate, especially given experiments showing that foreign-accented speech can be highly intelligible (Derwing & Munro, 1997; Derwing & Munro, 2005).

It is tempting therefore to redact the term 'native' in the descriptors presented here and use the term 'natural' instead (although this term can also be subject to questions such as "natural to whom"?). Such a move would resolve the issues outlined above and avoid much of the criticism levelled by researchers such as Rajadurai (2007), who argues that native speaking raters may regard non-native speakers as less intelligible because they have lower status pronunciation, rather than because it is objectively less easy to recognise as English. However, the reason nativeness is a suitable term to use in this scale is that the raters' judgements were made solely on the basis of their subjective interpretations of intelligibility. The evidence from these raters indicates that nativeness is an important way that native English speaking raters characterise intelligible delivery. It is also the way non-experienced raters, with no phonological training, describe their interpretation of these features of speech. The way intonation and rhythm promote intelligibility is consistent with them being similar to native speaker control over such features. Indeed, the findings presented in Chapter 5 where high-level speakers are routinely regarded as having native-like intonation tends to indicate that adult learners are perfectly capable of acquiring native control over rhythm and intonation.

It must be noted, though, that the current study employed English native-speaking raters only, and so an interesting extension to this project would be to examine whether non-native speaking

raters also refer to intonation and rhythm in terms of nativeness. It may be the case that they have a preference for referencing the ‘naturalness’ of the speech signal rather than the nativeness, which was also a term commonly applied in this research.

6.3.1 Reference to Existing Scales and Frameworks

An important step in examining the appropriateness of the suprasegmental descriptors presented in this chapter (henceforth referred to as the Suprasegmental Descriptors, for ease of reference) is to compare them to existing scales and frameworks. The purpose of this is twofold: to identify how well the outcome of this research fits with existing scholarship; and to identify whether this research achieves the goal of potentially complementing and supplementing the descriptors that are currently used in speaking tests. The suprasegmental features included in the following frameworks and scales are compared to the ones presented in this chapter:

1. the original phonological control grid of the CEFR (Council of Europe, 2001: 117)
2. Horner’s proposed revisions to original phonological control grid (Horner, 2010: 56)
3. the recently published revised CEFR phonological control grid (Council of Europe, 2017)
4. the Global Scale of English (GSE)
5. the Cambridge English Overall Speaking Scale (Cambridge English, 2016b: 83)
6. Pearson Test of English: Academic (Pearson, 2017a: 24)
7. TOEFL iBT Independent and Integrated Speaking Scales (ETS, 2014)

These scales and frameworks are reproduced in Appendices A and B. They represent a broad spectrum of the current approach to conceptualising pronunciation, although there are many more scales and frameworks that are not discussed here. For instance, the IELTS speaking band descriptors (IELTS, 2012: 19) are omitted because they make no explicit reference to suprasegmental features of speech.

CEFR Level B1

On the 2001 CEFR phonological control grid, intonation does not appear until B2 on the scale. This relates to a major motivation for this study, which is that suprasegmental features were not well elaborated on the CEFR. There is a similar issue in certain other frameworks and scales. For instance, the Cambridge English Overall Speaking Scale states at B1:

“is mostly intelligible, and has some control of phonological features at both utterance and word level”

(Cambridge English, 2016b: 83)

In the Suprasegmental Descriptors, at B1 focal stress is regarded as falling in unexpected places, intonation as being monotone, rhythm being fragmented, and lexical stress being incorrectly placed. Some parallels can be found in the 2017 CEFR phonological control grid, where the speaker’s delivery at B1 is characterised as intelligible despite “strong influence on stress, intonation and/or rhythm from other language(s)”, and speakers are able to “approximate intonation and stress at both utterance and word level” (Council of Europe, 2017: 134). The Suprasegmental Descriptors illustrate how that strong influence from an L1 might manifest itself, in terms of monotonicity, but on the whole it appears that the suprasegmental descriptors are rather harsher than the 2017 CEFR. They describe a speaker whose delivery is fragmented, and monotone, and who cannot adequately control focal stress.

At B1 the GSE tends to be consistent with the 2017 CEFR. Speakers are regarded as being able to “use basic stress and intonation to support meaning” (Pearson, 2016a: 8). The Suprasegmental Descriptors similarly describe such a speaker in rather more negative terms. Equally, Horner’s (2010) suggested alternative to the phonological control grid argues that at B1 “prominence is

sometimes used to effect” and “basic intonation patterns appear” (p56). This also tends to overestimate speaker control over these features at level B1, in contrast to the feedback from raters obtained in this study.

A possible explanation for the difference between existing scales and frameworks at B1 and the suprasegmental descriptors presented here is that negative terms have been used in the Suprasegmental Descriptors. It is difficult to discern whether there is a difference between “emphasis falls in unexpected places”, which appears at B1 on the Suprasegmental Descriptors, and “prominence is sometimes used to effect”, which appears on Horner’s scale. The PTE-A scales describe performance in negative terms in some instances, and it is possible to see parallels between the Suprasegmental Descriptors and the PTE-A descriptors. The PTE-A pronunciation scale at level 1 states: “stress may be placed in a non-English manner” (Pearson, 2017a: 24), which appears to be consistent with “emphasis falls in unexpected places”. Level 1 on the oral proficiency scale describes “irregular phrasing or sentence rhythm” and “staccato or syllabic timing”, which appear to correspond to the fragmented delivery that raters regarded as important at B1 on the suprasegmental scale. Unfortunately, the points on the PTE-A scales are not directly linked to the CEFR, so it is difficult to ascertain whether Level 1 on the oral fluency scale would approximately map to B1 on the CEFR. Similar parallels can be found between the Suprasegmental Descriptors and the TOEFL iBT scales, which also contain negative descriptors. A score of 1 on the independent speaking scale corresponds to delivery that is “choppy, fragmented”, and “consists of frequent pauses and hesitations”. And at level 2 “speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation or choppy rhythm/pacing”. This has clear parallels with B1 on the Suprasegmental Descriptors, where intonation required effort and fragmentation was a major characteristic of performances.

CEFR Level B2

The 2001 CEFR phonological control grid regards B2 speakers as having “clear, natural pronunciation and intonation” (Council of Europe, 2001: 117). This is not consistent with other scales and frameworks that tend to state that, although speakers are beginning to gain control over stress and intonation, they are not quite natural. The Cambridge Scale describes intonation as “generally appropriate” and focal stress as “generally accurately placed” at B2 (Cambridge English, 2016b: 83). The 2017 CEFR states that speakers can “employ prosodic features (e.g. stress, intonation, rhythm) to support the message”, and “can generally use appropriate intonation, place stress correctly” (Council of Europe, 2017: 134) at this level. And the GSE describes speakers as exhibiting “only minor hesitations and contributions, are intelligible using intonation and stress to convey meaning” (Pearson, 2016a: 9). Horner’s suggested descriptor follows the same pattern:

“Sounds and word stress are clearly intelligible. Features of linking appear.

Prominence used to effect. Basic intonation patterns are common.”

(Horner, 2010: 56)

B2 level on the Suprasegmental Descriptors contrasts somewhat with the descriptors of the other scales and frameworks. Although rhythm is not well represented in these frameworks it is mentioned in the 2017 CEFR where it is regarded as supporting the message. In the Suprasegmental Descriptors, effective rhythm does not emerge until B2+, indicating that it is an element of B2 but is certainly not well established at the beginning of the band. Equally, intonation is regarded as unnatural at B2 and speakers do not appear to gain control over it until B2+, where it is described as “effective”, with the speaker having “control over weak forms”, and “some appropriate focal stress”. It may be the case that these suprasegmental descriptors provide a more nuanced understanding of the difference between B1 and B2, and that there is not a clear separation between these levels, at least in terms of the suprasegmental aspect of pronunciation.

Again, the scales which permit negative descriptors, PTE-A and TOEFL iBT, do have some clear parallels to the Suprasegmental Descriptors. At level 3 on PTE-A speech “may be uneven” (Pearson, 2017a: 24), and as previously stated, TOEFL iBT describes performances as having “awkward intonation” and “choppy rhythm”.

Alternatively, B2 on the Suprasegmental Descriptors may not be well placed, and is perhaps more representative of B1+. This would be consistent with a leap in proficiency from B1 to B2 and would better reflect the original purpose of CEFR level B1 as representing basic operational proficiency and B2 as limited operational proficiency (North, 2005: 32). A calibration exercise, as discussed in Chapter 7, is necessary to establish the correct placement of the B2 and B2+ Suprasegmental Descriptors.

It appears to be the case then the Suprasegmental Descriptors presented here complement the descriptors developed for other frameworks. Specifically, they provide more detail as to the control over suprasegmental features exhibited at B2, but also effectively distinguish between B2 and B2+.

CEFR Level C1

The scales appear to converge more at the C1 level. Level C1 on the Suprasegmental Descriptors describes rhythm as progressing from limited choppiness, “it was a bit choppy, but still it wasn’t monotonous”, to rhythm becoming natural at C1+, “she pauses in just the right places, so she’s really got the rhythm of the language down to a tee”. Intonation progresses from periods of natural delivery, “very natural intonation”, to being native-like at C1+. Lexical stress progresses from “generally appropriate” to having “very little incorrect word stress”.

This is broadly consistent with the 2001 CEFR phonological control grid, where speakers “can vary intonation and place sentence stress correctly in order to express finer shades of meaning” (Council of Europe, 2001: 117), and it certainly reflects the “adequate operational proficiency” which C1 was originally intended to describe (Wilkins, 1978; North, 2005: 32). It is also consistent with Horner’s revision of the scale, where he regards speakers as being able to use intonation ‘to effect’ (Horner, 2010: 56), and with the Cambridge Scale, where intonation is described as “appropriate” and focal and lexical stress as “accurately placed” (Cambridge English, 2016b: 83). This is also consistent with the GSE, where a speaker is regarded as being able to “use stress and intonation to convey subtle or inferential meanings” (Pearson, 2016a: 9). Although the current study focuses on recognition rather than meaning, there is a clear parallel between natural and native-like control of intonation and its use to convey meaning. It also matches well with C1 on the 2017 CEFR, where speakers are characterised as being able to:

“produce smooth, intelligible spoken discourse with only occasional lapses in control of stress, rhythm and/or intonation, which do not affect intelligibility or effectiveness. Can vary intonation and place stress correctly in order to express precisely what he/she means to say.”

(Council of Europe, 2017: 134)

It is notable that scales tending to only use positive descriptors begin to more closely match the Suprasegmental Descriptors at C1, where raters started to describe performances in more positive terms, which in turn fed into more positive descriptors. The higher levels of scales that contain more negative descriptors also tend to align with the Suprasegmental Descriptors at C1. For instance, the TOEFL iBT independent speaking scale describes “minor difficulties with pronunciation or intonation”, which corresponds well to “L1 interference persists although the

speaker is now beginning to deliver periods of natural intonation” in the Suprasegmental Descriptors.

Summary

At levels B1 and B2 there is some inconsistency between the Suprasegmental Descriptors presented in this thesis and those included, or absent from, existing scales and frameworks. Specifically, the suprasegmental scale tends to appear somewhat harsher than the other scales at these levels. There are probably two reasons for these inconsistencies. One is the attachment to positive rating scale descriptors. In this study, negative descriptors were permitted for reasons discussed above in 6.2.2. In the scales where negative terms are avoided, gaps may exist due to the difficulty in formulating positive suprasegmental descriptors at low proficiency levels. This would also explain why the scales converge at C1 level, where the raters in this study used more positive language and the other scales are able to express more detail in positive terms.

The other possible reason for the difference between scales is the level of detail. Certain existing descriptors are so broad and lacking in detail, as discussed in the rationale to this study, that it is difficult to see how they relate to the Suprasegmental Descriptors described here. This may be described as necessary reductionism (Van Moere, 2013, cited in Harding, 2016: 13), but it tends to indicate that existing scales do not regard suprasegmental features to be salient, or consistent and transparent across levels. This study clearly indicates that suprasegmental features are salient and that they can be described in detail across levels. In addition to this, it should be reiterated that this study presents descriptors which are actually noticed by raters and can be described by them.

The rating scales and frameworks reviewed in this chapter rarely devote more than a single clause to a test taker's suprasegmental performance at each level. The Suprasegmental Descriptors presented here succeed in improving the degree of detail with which the suprasegmental elements of a test taker's performance are characterised. The possible applications of the descriptors and their implications are discussed in Chapter 7.

Chapter 7: Conclusion

The thesis concludes by reviewing the methodological approach and findings of the study, followed by a discussion of the limitations, and some potential future directions for research, including possible approaches to trialling the descriptors developed in Chapter 6. The chapter closes with a discussion of some potential applications for the suprasegmental rating scale descriptors and some concluding remarks that review the implications of this research and describe its contribution to the field.

A mixed-methods research design was employed to investigate the role of suprasegmental features in rater judgements of intelligibility. Scores on a 9-point intelligibility scale were collected and used to examine the extent to which raters were capable of judging suprasegmentally errant performances consistently. Scores were also used to corroborate the judgements made by experienced raters with those of typical listeners. Verbal reports were collected using a stimulated recall procedure. Raters described the features of each performance that influenced their judgements. This qualitative data was coded and categorised, and then used to identify which features of speech raters attended to when making their judgements. Good mixed-methods research does not simply address qualitative and quantitative data, but finds ways to integrate them (Hashemi & Babaii, 2013). This study has effectively integrated qualitative and quantitative data at the stages of data collection and analysis, culminating in the creation of an integrated display which was used to develop a series of suprasegmental rating scale descriptors.

The suprasegmental descriptors presented in Chapter 6 include greater detail on suprasegmental features of speech than existing rating scales. However, they also present a slightly different progression in some of these characteristics than other scales and frameworks. Specifically, at levels B1 and B2, the descriptors presented here are harsher than other such descriptors. As discussed in Chapter 6, this may be due to allowing negative descriptors, or to the difference in the level of detail. However, it is feasible that the descriptors, being based on perceptual recognition rather than broader understanding, are more sensitive to suprasegmental errors at these CEFR levels.

Harding (2016: 12) regards the primary difficulty in developing rating scale descriptors as coming from identifying what should be included in the scale, and ensuring that it is interpreted correctly by raters. The research presented here has attempted to overcome both these challenges by deriving the scale descriptors from the way in which raters described the performances.

7.1 Summary of Findings

Three research questions were posed at the outset of this project: RQ1 concerned the degree to which experienced raters agreed on how intelligible they found suprasegmentally errant speakers; RQ2 targeted the difference in scoring between experienced and non-experienced raters; and RQ3 related to which suprasegmental features of speech raters attended to when making their judgements. The overarching aim of this research was to establish the role of suprasegmental features of speech on pronunciation scales.

The findings will now be summarised in reference to each research question.

7.1.1 RQ1

- 1 Experienced raters are capable of consistently grading the intelligibility of speech samples that are strongly marked by non-standard suprasegmental features of speech.
- 2 Experienced raters are capable of doing so across three different types of feature: rhythm, intonation, and lexical stress.

Experienced raters were characterised as trained, experienced raters, with backgrounds in teaching English and learning foreign languages. They were asked to score 12 audio clips of test taker performances on a series of three 9-point ordered category items. The prompt for the items were non-technical questions designed to target comprehensibility, fluency, and intelligibility in turn. The first two items were included to encourage raters to clearly separate intelligibility from fluency and comprehensibility. The within-group analysis of experienced rater responses to the intelligibility item indicated a close consistency of scores. Given the common background of the raters, it might be expected that they would have a common perception of the intelligibility of the performance. However, the audio stimuli were chosen because they reflected three broad categories of non-standard suprasegmental features of speech: intonation, rhythm, and lexical stress. Therefore, it appears to be the case that raters are sensitive to these features and consistent in the way they regard them as influencing intelligibility.

7.1.2 RQ2

- 1 Non-experienced raters are also capable of making consistent judgements of the intelligibility of speech consisting of non-standard suprasegmental features.
- 2 This consistency is exhibited across different types of suprasegmental features, specifically characteristics of rhythm and intonation.
- 3 Non-experienced raters appear more severe and more reliable than experienced raters.

- 4 Experienced raters reported sensitivity to their own variability in terms of accent familiarity.

Non-experienced raters were characterised as those who lacked training, rating experience, and experience of teaching English and learning foreign languages. As such, they represent the type of listener a test taker might encounter in a real-world context. They listened to the same audio stimuli as the experienced raters and responded to the same questions. These raters were consistent with one another across the suprasegmental features represented in the audio stimuli. The between-groups analysis indicated that they were somewhat more severe than the experienced raters, probably due to their limited exposure to foreign-accented English. Notably though, experienced raters described an awareness of the non-experienced listener when making their decisions and may well have been moderating their scores to accommodate the typical listener. On this basis, it is possible to conclude that, in the language testing context, the scoring of suprasegmental features by experienced raters exhibits cognitive validity in the sense that their judgements are broadly consistent with the judgements of typical listeners.

7.1.3 RQ3

- 1 Experienced raters referred to characteristics of rhythm, intonation, and lexical stress when describing how they made their judgments of intelligibility. There was broad consistency within the experienced rater group as to the influence of these features.
- 2 Non-experienced raters attended to characteristics of rhythm and intonation when describing how they made their judgements. There was consistency among the non-experienced raters as to the role of these features to intelligibility.
- 3 Raters used similar terms to refer to suprasegmental variation: experienced raters were more nuanced and technical, whereas non-experienced raters were broader and less technical.

- 4 The three broad classes of suprasegmental features were sufficiently well distinguished, and the terms raters used were adequately mutually comprehensible, so that a series of suprasegmental rating scale descriptors could be developed.

Raters had a common perception of both the intelligibility of the speakers they were presented with and the specific suprasegmental features of speech that promoted or detracted from their performances. The inclusion of a range of experienced raters allows this finding to be generalised to the broad community of native-English-speaking assessors. The inclusion of non-experienced raters demonstrates that the findings are not simply reserved for language specialists, but can be expected to be representative of communicative interactions with typical listeners. It is possible to conclude that the shared terms used by raters are suitable for rating scale descriptors.

7.2 Applications

There are several potential applications for the suprasegmental descriptors presented in Chapter 6. These relate to language assessment stakeholders, including test developers, raters, test takers, language teachers, and score users.

Test Developers

Given the limited detail afforded to suprasegmental features on certain rating scales and frameworks, there is a risk of such features becoming hidden aspects of the assessment criteria. At worst, raters may be employing their own idiosyncratic and inconsistent approach to judging this facet of pronunciation, which causes bias and therefore damages the fairness of the exam.

By developing a series of rating scale descriptors that are grounded in the way raters and typical listeners interpret test taker speech signals, this study presents practical guidance to rating scale

Conclusion

developers on an appropriate way to incorporate suprasegmental features into pronunciation assessments. Dealing with suprasegmental features in more detail provides better direction to raters and more explicitly articulates the pronunciation construct, thus improving exam validity.

Raters

The empirical evidence gathered in this study, regarding the ways in which raters responded to suprasegmental features of speech, would be useful in rater training for assessing pronunciation. Training based on the comments and scores collected for this study, and the suprasegmental descriptors developed in Chapter 6, would boost raters' confidence in rating pronunciation, as well as enhancing the scoring validity of the tests that raters would go on to judge.

Test takers

The Suprasegmental Descriptors developed in Chapter 6 could be used by prospective test takers as a self-assessment or peer-assessment tool. They would enable test takers to be aware of the standards that raters apply when judging assessment, and to compare their current performance to that standard. The descriptors were constructed predominantly out of non-technical language, meaning with minor amendments they would be a powerful tool for learners to use to identify the strengths and weaknesses in their suprasegmental performance. Additionally, they could be employed in the language classroom to contribute to learning-oriented assessment and during general L2 English courses.

Language Teachers and Material Writers

Suprasegmental features of speech are often described as somewhat neglected in the pronunciation literature (e.g. Isaacs, 2008). This is beginning to change, for example, with the

development of teaching resources designed to target the control of suprasegmental features by learners of specific first languages (Frost & Picavet, 2014; Frost & O'Donnell, in press). Given the relative lack of research in this area, it is worth identifying to what extent the findings presented here can be expected to aid teaching of pronunciation and development of syllabi.

The suggestions below, which follow directly from the findings of this research, may act as guidance for teachers or syllabus designers who are involved in exam preparation for test takers at the B1 – C1 level.

Rhythm Attention to specific rhythmic timing does not appear to be necessary at level B1 – C1. Instead, learners should focus on overcoming the hesitations that cause their delivery to be fragmented. This is likely to require a holistic approach which targets linguistic, cognitive, and affective factors. It may be facilitated by identifying where 'natural' pauses should fall, and the role of weak forms in connected speech.

Intonation Learners should emulate proficient speakers to cultivate a natural or native-like delivery. Specific features that learners can focus on to facilitate this are appropriate focal stress placement and reducing monotonicity of delivery. Control over specific pitch patterns is less important to intelligibility, except in avoiding delivering excessive rising tones.

Lexical Stress Teachers may want to focus on the correct pronunciation of certain key words that are commonly mis-stressed and known to interfere with intelligibility. Not all lexical stress errors are equally damaging to intelligibility, and certain lexical stress errors are tolerable, meaning that spending time teaching correct pronunciation of lexical stress errors that do not severely influence intelligibility may be inefficient.

Score Users

Score users include: companies that want to ensure the people they employ have a certain proficiency in English; universities that need to ascertain whether new students will be capable of studying in an English; and governments that require new migrants to speak English in order to demonstrate their ability to survive and integrate into a new country. Each of these users has a stake in the quality of the speaking tests that their applicants take. There are primarily two concerns in this regard, firstly that the entry standards that organisations set are reflective of the English proficiency they require of applications. For instance, students entering a British university may be required to deliver presentations to a group of listeners, and as such they require the entrant to achieve a minimal level of intelligibility in spoken English. The minimum score that applicants are required to achieve on the tests must therefore reflect this minimal level of intelligibility. The second and final concern is that the test is reliable, that measurement error is minimised and the level of English proficiency defined by the exam is consistent from test taker to test taker. Without this element of reproducibility, the scores lose their meaning.

By providing more insight into the role of suprasegmental features into the assessment of pronunciation, the risk of these two concerns is mitigated. More concrete pronunciation criteria lead to more precise measurement. This more precise measurement leads to there being less scope for raters to engage their own idiosyncratic judgments of the performances, which in turn leads to more consistent judgments. Furthermore, the use of non-experienced raters in this study provides a link from professional raters to the typical listeners that a test taker may encounter in their score use context. In summary, the implications of exams employing some variant of the descriptors presented here is greater precision of measurement for score users, leading to better decisions in contexts such as university admissions.

7.3 Limitations of the Study and Future Directions

There are several limitations to the study that have not already been discussed, one of which relates to task type and mode. In order to obtain test takers across the required ability range, it was necessary to use audio clips from three different exams: PET, FCE, and CAE. The material used in this research was the long-turn monologue from each exam. There is a difference in the prompts used in each of these exams. For example, in PET candidates are asked to “tell us what you can see in the photograph” (Cambridge English, 2016c), whereas on FCE and CAE they are asked to compare some photographs and then respond to a specific point such as “how important is it to help people in these situations” (Cambridge English, 2016a). There is evidence that task type influences candidate performance (Skehan & Foster, 1997), but it is not clear if or how these different instructions would influence a candidate's ability to demonstrate their pronunciation competence, and specifically to express a range of suprasegmental features. It may be the case that there is more scope for suprasegmental variation in delivery when comparing two pictures than when describing a single image. In the current study, this was ameliorated by the active selection of audio clips on the basis of the suprasegmental feature they exhibited. Nonetheless, the examination medium is a particularly interesting route for further research. As Knoch (2016: 63-4), states there is a clear need for more research into task effects in speaking assessments, and this is also the case in terms of how effectively such tasks provide speakers with scope to express a range of suprasegmental features of speech. In future, a study that employs approximately the same methodology as this one but differentiates on the basis of task would help to ascertain what influence the type of instructions have on the resulting test taker deliveries.

This study used long-turn monologues for tasks, where test takers were asked to talk about images. Dialogues or other modes of interaction employ suprasegmental features, so it would be interesting to identify whether the rating scale descriptors presented in Chapter 6 are suitable for

use in assessing dialogues. It is probable that such a scale would include some focus on the intonation associated with ceding the speaker's turn and other such communicative tasks which require an interactive interlocutor. Indeed, the choice of using monologue rather than dialogue, although well-rationalised for practical reasons, does limit the opportunity for much of the intonation associated with turn taking to be expressed. Certain characteristics of suprasegmental delivery that are unique to interactions are likely to be underrepresented in this study. This does not necessarily represent an issue of content validity, because many people learn English for delivering lectures and presentation, for instance, which are primarily monologues. Indeed, some research into focal stress has taken the academic lecture as a context, since appropriate placement of focal stress is regarded as promoting listener motivation (Pickering, 2001).

One of the main observations of this study is that there was a small variation in the ratings made by experienced and non-experienced raters and it was proposed that this variation was caused by experienced raters having broader experience of listening to non-native speakers. It is important, as a next step, to control more closely the linguistic experience of the raters to identify the extent to which this influences judgements. In order to control for the multitude of variables at play in a spoken interaction, the decision was made to limit the raters to native-English speakers. Certainly, this represents a potential deviation from the communicative situations that a test taker may encounter, for instance when speaking English on a business trip; furthermore, some examining bodies allow non-native raters to assess speaking. As such, there might be scope for a second study investigating the consistencies between native and non-native English speaking raters. A further study examining the links between specific linguistic backgrounds of raters and their perception of suprasegmental features of speech, similar to the work of Winke et al. (2013), would establish whether the descriptors presented in this study are suitable for the general population of raters. Ultimately, without broader validation, the scale descriptors presented here

are primarily suitable to be used by English native speaking raters for the examination of monologues.

Although rater linguistic background was controlled as much as feasible, the L1 of the test takers was not restricted. Indeed, a broad range of speakers was purposefully sampled in order to ensure a broad range of suprasegmental anomalies were present. It would be interesting to understand how effectively the descriptors presented in Chapter 6 can be employed to assess learners from specific speaking communities. For instance, Frost and O'Donnell (in press), have developed such a scale specifically for use in assessing French speakers of English.

Two final areas of further enquiry are the scale criteria and test taker ability level. The descriptors presented in this thesis target solely the ability of listeners to recognise the speech signal. As such, they provide good insight into the initial level of processing that test takers are required to undertake and specifically target pronunciation, which is highly relevant to the testing context. But nonetheless, without reference to comprehensibility, they represent an incomplete picture as to the role of suprasegmental features in oral proficiency overall. An interesting expansion of this project would be to examine the interaction between suprasegmental features of speech and criteria such as comprehensibility, fluency, and accent.

Equally, a final area for potential future research is examining the influential suprasegmental errors that take place at CEFR levels above C1. This research concerned levels B1 – C1, and although this leaves the levels below B1 unexamined, it seems unlikely that there will be major significant errors at these levels. This is because, suprasegmental control does not begin to emerge until the range B1 – C1 according to the descriptors developed here, as well as those currently in use. However, it would be interesting to see how performance emerges post-C1. As speech becomes much more natural, raters may begin to concern themselves more with specifics

of the rhythmic timing, and may consistently attend to complex tone misplacement. Alternatively, it may be the case that suprasegmental features begin to subside as a risk to intelligibility at this level.

7.3.1 Trialling Suprasegmental Descriptors

The primary way in which this study could be expanded is by trialling the descriptors that were presented in Chapter 6. The descriptors presented here have been closely derived from the data collected. They have not been trialled, however, and the next step in this line of enquiry would be to assess how raters use them. This would confirm both the usability of the descriptors and allow them to be scaled appropriately. What follows is a proposal for trialling the Suprasegmental Descriptors.

There are two characteristics of the descriptors that trialling would examine: how easily and consistently raters apply them; and their placement relative both to each other and to the CEFR. This would require a mixed-methods approach which collects raters' perceptions of the usability of the descriptors, and uses quantitative tools to evaluate their efficacy. The quantitative aspect of a trial of these descriptors would require a minimum of 250 raters to use the scale to judge a series of test takers. The Rasch Rating Scale model would then be employed, as it was in this study, to examine the way in which raters employ the descriptors. This measurement-driven approach would make it possible to address the quality of the descriptors in terms of whether they progress monotonically, and are applied in an appropriately predictable way. Anchor descriptors with known CEFR values would also be employed. By calibrating the Suprasegmental Descriptors alongside the anchor descriptors, it would be possible to scale the Suprasegmental Descriptors to the CEFR, and to establish how effectively they mirror the estimated CEFR levels presented in

Chapter 6. An example of this approach is the descriptors used in the Global Scale of English, which are calibrated in reference to North (2000) (de Jong & Benigno, 2016).

The quantitative approach would provide insight into how effectively raters use the descriptors in reference to how well they fit the Rasch model. However, a full understanding of their usability requires rich qualitative data examining rater approaches to using the rating scale descriptors. Harding (2016) reports on a suitable methodology for examining descriptor usability. He asked raters to judge a series of speech samples using the Common European Framework of Reference for languages (Council of Europe, 2001), and then discussed their experience of applying the scale in a focus group format. The outcome was an understanding of issues related to the clarity of the descriptors, their conciseness and intuitiveness, as well as their theoretical suitability.

In practical terms, if the quantitative aspect of such a study employed 250 raters, a fraction of these could be invited to take part in one or more focus groups. The descriptors presented here were devised on the basis of rater responses, so it is highly likely that raters would be able to consistently employ the descriptors across levels, but nonetheless the application of Harding's approach would provide concrete empirical evidence of their usability, and would provide insight into how they could be improved.

7.4 Concluding Remarks

This study has contributed to the literature in several ways. Primarily, it provides empirical evidence for the suprasegmental features of speech that raters attend to while making their judgments of intelligibility. In addition, it illustrates the degree to which raters are consistent in which features they find to be influential. This suggests that raters share a common hierarchy of the suprasegmental features of speech that they regard as influential to intelligibility. Importantly,

Conclusion

this hierarchy has also been found to be broadly consistent with the way typical listeners interpret these features. In other words, the contribution this project makes to the field is to identify the specific suprasegmental features of speech that consistently influence rater judgements in reference to intelligibility, and therefore to identify the most suitable features for consideration as rating scale criteria.

Understanding how raters interpret pronunciation is important because pronunciation, as it appears on rating scales, is not necessarily the *de facto* construct insofar as raters apply it during examinations. This project brings more precision to defining the actual construct that raters regard as important to pronunciation in suprasegmental terms. As such, it provides a critical step in exam validation by improving the provision of construct definitions (Weir, 2005: 18).

This study broadens the scope of current pronunciation research by investigating perceived intelligibility as an oral proficiency criterion. This may well be the first time the relationship between suprasegmental features, as raters perceive them, and intelligibility, as measured on a rating scale, has been investigated. More concretely, this research provides test developers with a better idea of the pronunciation construct, and an indication of the level at which raters interpret the speech signal; it presents concrete descriptors which language assessors can employ to judge the intelligibility of pronunciation at levels B1 – C1 on the CEFR; and it gives teachers and learners an indication of the suprasegmental characteristics of their deliveries that are likely to be influential to test scores when taking an exam.

References

- Abercrombie, D. (1949). Teaching pronunciation. *ELT Journal*, 3(5), 113-122.
- Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.
- ACTFL. (2012). ACTFL proficiency guidelines. Retrieved on 13-04-2016 from http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf
- Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 98(23).
- Albrechtsen, D., Henriksen, B., & Faerch, C. (1980). Native speaker reactions to learners' spoken interlanguage 1. *Language Learning*, 30(2), 365-396.
- Alderson, J. C. (1991) Bands and scores. In Alderson, C. and North, B. (eds.) *Language testing in the 1990s*. MEP: British Council.
- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Allen, G. D. (1975). Speech rhythms: Its relation to performance universals and articulatory timing. *Journal of Phonetics*, 3(2), 75-86.
- ALTE. (2002). The ALTE can do project. Retrieved on 15-04-2016 from <http://www.cambridgeenglish.org/images/28906-alte-can-do-document.pdf>
- 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(4), 529-555.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573.
- Ang-Aw, H. T., & Goh, C. C. M. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC Journal*, 42(1), 31-51.
- Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, 40(3), 351-373.

References

- Arvaniti, A., & Rodriguez, T. (2013). The role of rhythm class, speaking rate, and F0 in language discrimination. *Laboratory Phonology*, 4(1), 7-38.
- Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, 66(1-2), 46-63.
- Ashby, M., & Maidment, J. (2005). *Introducing phonetic science*. Cambridge: Cambridge University Press.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
- Ballard, L., & Winke, P. (2016). Students' attitudes towards English teachers' accents: The interplay of accent familiarity, comprehensibility, intelligibility, perceived native speaker status, and acceptability as a teacher. In T. Isaacs, & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 121-140). Bristol: Multilingual Matters.
- Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 7(1), 54-74.
- Barnwell, D. (1989). 'Naive' native speakers and judgements of oral proficiency in Spanish. *Language Testing*, 6(2), 152-163.
- Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, 2(1), 49-58.
- Bazeley, P. (2010). Computer assisted integration of mixed methods data sources and analyses. In A. Tashakkori, & C. Teddlie (Eds.), *SAGE handbook of mixed methods in social & behavioral research* (2nd ed., pp. 431-467). Thousand Oaks, CA: Sage.
- Beckman, M. E. (1986). *Stress and non-stress accent*. Dordrecht: Foris.
- Beckman, M. E. (1996). The parsing of prosody. *Language and Cognitive Processes*, 11(1-2), 17-68.
- Beckman, M. E. (1992). Evidence for speech rhythms across languages. In Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure* (pp. 457-463). Tokyo: IOS Press.
- Beebe, L. M. (1980). Sociolinguistic variation and style shifting in second language acquisition. *Language Learning*, 30(2), 433-445.

- Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, 114(3), 1600-1610.
- Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27(3), 355-377.
- Birch, S., & Clifton, C. (1995). Focus, accent, and argument structure: Effects on language comprehension. *Language and Speech*, 38(4), 365-391.
- Bolinger, D. L. (1958). A theory of pitch accent in English. *Word-Journal of the International Linguistic Association*, 14(2-3), 1-149.
- Bolinger, D. L. M. (1972). *Intonation: Selected readings*. Harmondsworth: Penguin.
- Bond, T., & Fox, C. M. (2015). *Applying the rasch model: Fundamental measurement in the human sciences*. London: Routledge.
- Bond, Z. S. (2008). Slips of the ear. In D. B. Pisoni, & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 290-310). Oxford: Blackwell.
- Bond, Z. S., & Small, L. H. (1983). Voicing, vowel, and stress mispronunciations in continuous speech. *Attention, Perception, & Psychophysics*, 34(5), 470-474.
- 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(04), 447-465.
- Brazil, D. (1997). *The communicative value of intonation in English*. Cambridge: Cambridge University Press.
- Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing*, 29(1), 91-108.
- Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Logan, Utah: Utah State University Press.
- Brown, A., & Taylor, L. (2006). A worldwide survey of examiners' views and experience of the revised IELTS speaking test. *Research Notes*, 26, 14-18.
- Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. *International English Language Testing System (IELTS) Research Reports*, 3, 49-84.
- Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *ETS Research Report Series*, 2005(1).

References

- Browne, K. (2016). *Raters' accent-familiarity levels and their effects on pronunciation scores and intelligibility on high-stakes English tests*. (Unpublished PhD). University of Leicester,
- Browne, K., & Fulcher, G. (2016). Pronunciation and intelligibility in assessing spoken fluency. In T. Isaacs, & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 37-53). Bristol: Multilingual Matters.
- Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.
- Bybee, J. (2010). *Language, usage and cognition* Cambridge: Cambridge University Press.
- Byrne, D. (1986). *Teaching oral English*. London: Longman.
- Callaway, D. R. (1977). Accent and the evaluation of ESL oral proficiency: Occasional papers on linguistics, no. 1. *Proceedings of the International Conference on Frontiers in Language Proficiency and Dominance Testing*, Southern Illinois University, Carbondale, Illinois. 163-177.
- Cambridge Assessment. (2017a). Become a Cambridge English examiner. Accessed on 15-06-2017 from <http://www.cambridgeenglish.org/teaching-english/become-a-cambridge-english-examiner/>
- Cambridge Assessment. (2017b). Cambridge English: First sample paper. Retrieved on 23-05-2017 from <http://www.cambridgeenglish.org/exams-and-tests/first/exam-format/>
- Cambridge Assessment. (2017c). Who can take Cambridge English exams? Accessed on 01-07-2017 from <https://support.cambridgeenglish.org/hc/en-gb/articles/202838466-Who-can-take-Cambridge-English-Exams->
- Cambridge English. (2015). *The Cambridge English scale explained*. Cambridge English Language Assessment. Retrieved on 13-10-2016 from <http://www.cambridgeenglish.org/images/177867-the-methodology-behind-the-cambridge-english-scale.pdf>
- Cambridge English. (2016a). Cambridge English: First (FCE), preparation, free: Paper-based sample test. Retrieved on 10-11-2016 from <http://www.cambridgeenglish.org/exams/first/preparation/>
- Cambridge English. (2016b). Cambridge English: First handbook for teachers. Retrieved on 27-09-2016 from <http://www.cambridgeenglish.org/images/167791-cambridge-english-first-handbook.pdf>
- Cambridge English. (2016c). Cambridge English: Preliminary (PET), preparation, free: Paper-based sample test. Retrieved on 27-09-2016 from <http://www.cambridgeenglish.org/exams/preliminary/preparation/>

- Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, 28(2), 201-219.
- Caspers, J. (2010). The influence of erroneous stress position and segmental errors on intelligibility, comprehensibility and foreign accent in dutch as a second language. *Linguistics in the Netherlands*, 27(1), 17-29.
- Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching pronunciation hardback with audio cds (2): A course book and reference guide*. Cambridge: Cambridge University Press.
- Chadwick, E. (1864). The museum, a quarterly magazine of education, literature and science. vol. II. *Journal of the Statistical Society of London*, 27(2), 261-266.
- Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.
- Chalhoub-Deville, M. (1995). A contextualized approach to describing oral language proficiency. *Language Learning*, 45(2), 251-281.
- Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes*, 24(3), 383-391.
- Chalhoub-Deville, M., & Fulcher, G. (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals*, 36, 498-506.
- Charmaz, K. (2014). *Constructing grounded theory* (2nd ed.). London: Sage.
- Chen, H. C., & Wang, Q. (2016). The effects of Chinese learners' English acoustic-prosodic patterns on listeners' attitudinal judgments. *The Southeast Asian Journal of English Language Studies*, 22(2), 91-108.
- Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, 6(3), 271-315.
- Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, 116(6), 3647-3658.
- Classe, A. (1939). *The rhythm of English prose*. London: Blackwell.
- Coates, J. (1993). *Women, men and language*. London: Longman.
- Cooper, N., Cutler, A., & Wales, R. (2002). Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners. *Language and Speech*, 45(3), 207-228.

References

- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Council of Europe. (2017). *Common European framework of reference for languages: Learning, teaching, assessment. companion volume with new descriptors*. Strasbourg: Council of Europe.
- Cowie, R., Douglas-Cowie, E., & Romano, A. (1999). Changing emotional tone in dialogue and its prosodic correlates. *ESCA Tutorial and Research Workshop (ETRW) on Dialogue and Prosody*, Veldhoven, The Netherlands.
- Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). 'FEELTRACE': An instrument for recording perceived emotion in real time. *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, NI.
- Creswell, J. W. (2014). *Educational research: Planning, conducting, and evaluating quantitative* (4th ed.). Harlow: Pearson Education.
- Crisp, V. (2008). The validity of using verbal protocol analysis to investigate the processes involved in examination marking. *Research in Education*, 79(1), 1-12.
- Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *The Modern Language Journal*, 99(1), 80-95.
- Cruttenden, A. (1997). *Intonation*. Cambridge: Cambridge University Press.
- Crystal, D. (1969). *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.
- Crystal, D. (2011). *Dictionary of linguistics and phonetics*. London: John Wiley & Sons.
- Crystal, D., & Quirk, R. (1964). *Systems of prosodic and paralinguistic features in English*. Hague: Mouton.
- Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, 86(1), 67-96.
- Cutler, A., & Clifton Jr, C. (1984). The use of prosodic information in word recognition. In Bouma, H., & Bouwhuis, D. G. (Ed.), *Attention and performance 10: Control of language processes* (pp. 183-196). Hove: Psychology Press.
- Cutler, A., & Clifton, C. (1999). Comprehending spoken language: A blueprint of the listener. In C. M. Brown, & P. Hagoort (Eds.), *The neurocognition of language* (pp. 123-166). Oxford: Oxford University Press.

- Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40*(2), 141-201.
- Cutler, A. (1986). Forbear is a homophone: Lexical prosody does not constrain lexical access. *Language and Speech, 29*(3), 201-220.
- Cutler, A. (2005). Lexical stress. In D. Pisoni, & R. Remez (Eds.), *The handbook of speech perception* (pp. 264-289). Oxford: Blackwell.
- Cutler, A. (2015). Lexical stress in English pronunciation. In M. Reed, & J. M. Levis (Eds.), *The handbook of English pronunciation* (pp. 106-124). Oxford: Wiley Blackwell.
- Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language, 2*(3-4), 133-142.
- Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14*(1), 113.
- Cutler, A., & Pasveer, D. (2006). Explaining cross-linguistic differences in effects of lexical stress on spoken-word recognition. *3rd International Conference on Speech Prosody, May 2-5, Dresden*
- Dalton, C., & Seidlhofer, B. (1994). *Pronunciation*. Oxford: Oxford University Press.
- Dang, T. N. D. (2013). Impact of rhythm on Vietnamese adult EFL learners intelligibility in term of mid-level tone. *International Journal of Applied Linguistics & English Literature, 2*(4), 98-109.
- Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics, 11*, 51-62.
- Davies, Alan (2011) Does Language Testing Need the Native Speaker? *Language Assessment Quarterly, 8*(3), 291-308.
- Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135.
- de Jong, J., & Benigno, V. (2016). The CEFR in higher education: Developing descriptors of academic English. *Language Testing Forum, Nov 26-27*. University of Reading.
- Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(01), 1-16.
- Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A Research-Based approach. *Tesol Quarterly, 39*(3), 379-397.

References

- Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, 48(3), 393-410.
- Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, 54(4), 655-679.
- Díez, F. G., Dellwo, V., Gavaldà, N., & Rosen, S. (2008). The development of measurable speech rhythm during second language acquisition. *The Journal of the Acoustical Society of America*, 123(5), 3886-3886.
- Douglas, D., & Chapelle, C. (1993). A new decade of language testing research. In D. Douglas, & L. Selinker (Eds.), *Performance on a general versus a field-specific test of speaking proficiency by international teaching assistants* (pp. 235-256). Alexandria, VA: TESOL Publications.
- Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing*, 11(2), 125-144.
- Douglas, D., & Selinker, L. (1992). Analyzing oral proficiency test performance in general and specific purpose contexts. *System*, 20(3), 317-328.
- Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, 26(3), 423-443.
- Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet rasch analysis. *Language Assessment Quarterly: An International Journal*, 2(3), 197-221.
- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155-185.
- Eckes, T. (2011). *Introduction to many-facet rasch measurement*. Frankfurt: Peter Lang.
- Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing*, 10(3), 235-254.
- Ericsson, K. A., & Simon, H. A. (1985). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
- ETS. (2014). TOEFL speaking rubrics. Retrieved on 27-09-2017 from http://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf
- Faure, G., Hirst, D. J., & Chafcouloff, M. (1980). Rhythm in English: Isochronism, pitch and perceived stress. In L. Waugh, & C. van Schooneveld (Eds.), *The melody of language* (pp. 71-79). Baltimore: University Park Press.

- Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, 37(3), 313-326.
- Field, J. (forthcoming). Second language listening: Current ideas, current issues. In *Cambridge handbook of second language learning*. Cambridge: Cambridge University Press.
- Field, J. (2003). *Psycholinguistics: A resource book for students*. Hove: Psychology Press.
- Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, 39(3), 399-423.
- Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Studies in language testing 30: Examining speaking* (pp. 65-111). Cambridge: Cambridge University Press.
- Fisher, W. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, 21(1), 1095.
- Fox, A. (2000). *Prosodic features and prosodic structure: The phonology of suprasegmentals*. Oxford: Oxford University Press.
- Frost, D., & O'Donnell, J. (in press). *Pronunciation of English as a foreign or second language*. evaluating the essentials, the place of prosody in oral production. In J. Volin (Ed.), *The pronunciation of English as a foreign or second language*. Cambridge: Cambridge University Press.
- Frost, D., & O'Donnell, J. (2015). Evaluating the essentials: The place of prosody in oral production. *ELIP4 4 Th International Conference on English Pronunciation: Issues & Practices*, May 21-23, Prague. 34-36.
- Frost, D., & Picavet, F. (2014). Putting prosody First—Some practical solutions to a perennial problem: The innovalangues project. *Research in Language*, 12(3), 233-243.
- Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America*, 27(4), 765-768.
- Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, 1(2), 126-152.
- Fujii, S., & Wan, C. Y. (2014). The role of rhythm in speech and language rehabilitation: The SEP hypothesis. *Frontiers in Human Neuroscience*, 8, 777.
- Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign language*. (Unpublished PhD). University of Lancaster.

References

- Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13(2), 208-238.
- Fulcher, G. (2003). *Testing second language speaking*. Harlow: Pearson Education.
- Fulcher, G. (2004). Deluded by artifices? The common European framework and harmonization. *Language Assessment Quarterly: An International Journal*, 1(4), 253-266.
- Fulcher, G. (2015). Assessing second language speaking. *Language Teaching*, 48(02), 198-216.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London: Routledge.
- Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, 28(1), 5-29.
- Galaczi, E., Lim, G., & Khabbazzbashi, N. (2012). Descriptor salience and clarity in rating scale development and evaluation. *Language Testing Forum*, Nov 16-18, Bristol, UK.
- Galaczi, E., & French, A. (2011). Context validity. In L. Taylor (Ed.), *Studies in language testing 30: Examining speaking* (pp. 112-170). Cambridge: Cambridge University Press.
- Galaczi, E. D., French, A., Hubbard, C. & Green, A. (2011). Developing assessment scales for large-scale speaking tests: a multiple-method approach. *Assessment in Education: Principles, Policy & Practice*, 18(3): 217–237.
- Galaczi, E., Post, B., Li, A., Barker, F., & Schmidt, E. (2016). Assessing second language pronunciation: Distinguishing features of rhythm in learner speech at different proficiency levels. In T. Isaacs, & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 157-182). Bristol: Multilingual Matters.
- Galaczi, E., Post, B., Li, A., & Graham, C. (2011). Measuring L2 English phonological proficiency: Implications for language assessment. *Proceedings of the 44th Annual Meeting of the British Association for Applied Linguistics, the Impact of Applied Linguistics*, 67-72. London: Scitsiugnil Press.
- Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. London: Routledge.
- Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, 34(1), 65-87.
- Giegerich, H. J. (1992). *English phonology: An introduction*. Cambridge: Cambridge University Press.

- Gilbert, S., & Staub, G. (2014). Examiner confidence survey: An investigation into speaking examiners' confidence in the accuracy of the assessments they make. *Cambridge English: Research Notes*, 57, 50-59.
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine.
- Goldinger, S. D. (1997). Words and voices—perception and production in an episodic lexicon. In K. Johnson, & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 33-66). San Diego, CA: Academic Press.
- Grabe, E., Kochanski, G., & Coleman, J. (2005). The intonation of native accent varieties in the British isles—potential for miscommunication? In K. Dziubalska-Kořaczyk, & J. Przedlacka (Eds.), *English pronunciation models: A changing scene* (pp. 311-337). Bern: Peter Lang.
- Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology*, 7(515-546)
- Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press.
- Gui, M. (2012). Exploring differences between Chinese and American EFL teachers' evaluations of speech performance. *Language Assessment Quarterly*, 9(2), 186-203.
- Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press.
- Gut, U. (2012). Rhythm in L2 speech. *Speech and Language Technology*, 14/15, 83-94.
- Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, 38(2), 201-223.
- Halliday, M. (1970). *A course in spoken English: Intonation*. Oxford: Oxford University Press.
- Halliday, M. A. (1963). The tones of English. *Archivum Linguisticum*, 15(1), 1-28.
- Halliday, M. A. K. (1967). *Intonation and grammar in British English*. The Hague: Mouton.
- Hammond, M. (1999). *The phonology of English: A prosodic optimality-theoretic approach: A prosodic optimality-theoretic approach*. Oxford: Oxford University Press.
- Han, Q. (2016). Rater cognition in L2 speaking assessment: A review of the literature. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, 16(1), 1-24.

References

- Harding, L. (2013). Pronunciation assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ: Wiley-Blackwell.
- Harding, L. (2016). What do raters need in a pronunciation scale?: The users' view. In T. Isaacs, & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 12-34). Bristol: Multilingual Matters.
- Harding, L. (2017). Validity in pronunciation assessment. In O. Kang, & A. Ginther (Eds.), *Assessment in second language pronunciation*. London: Routledge.
- Hashemi, M. R., & Babaii, E. (2013). Mixed methods research: Toward new research designs in applied linguistics. *The Modern Language Journal*, 97(4), 828-852.
- Hayes-Harb, R., & Hacking, J. F. (2015). Beyond rating data: What do listeners believe underlies their accentedness judgments? *Journal of Second Language Pronunciation*, 1(1), 43-64.
- Hill, D., Jassem, W., & Witten, I. (1979). A statistical approach to the problem of isochrony in spoken British English. In H. Hollien, & P. Hollien (Eds.), *Current issues in linguistic theory 9: Current issues in the phonetic sciences* (pp. 285-294). Amsterdam: John Benjamins.
- Hill, K. (1996). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. *Melbourne Papers in Language Testing*, 5(2), 29-50.
- Hitzman, D. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review*, 93(4), 411-428.
- Hoffman, R. R., Shadbolt, N. R., Burton, A. M., & Klein, G. (1995). Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes*, 62(2), 129-158.
- Horner, D. (2010). A critical look at the CEFR "phonological control" grid. *IATEFL Testing, Evaluation and Assessment Special Interest Group (TEA SIG); EALTA Conference, Barcelona, Spain*. 50-57.
- Horner, D. (2013). Towards a new phonological control grid. In E. D. Galaczi, & C. J. Weir (Eds.), *Exploring language frameworks: Proceedings of the ALTE Kraków conference, July 2011* (pp. 227). Cambridge: Cambridge University Press.
- Horner, D. (2014). And what about testing pronunciation? A critical look at the CEFR pronunciation grid and a proposal for improvement. In R. van den Doel, & L. Rupp (Eds.), *Pronunciation matters accents of English in the Netherlands and elsewhere* (pp. 109-124)

- Hsieh, C. (2011). Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. *Spain Fellow Working Papers in Second Or Foreign Language Assessment*, 9, 47-74.
- Hubbard, C., Gilbert, S., & Pidcock, J. (2006). Assessment processes in speaking tests: A pilot verbal protocol study. *Research Notes*, 24, 14-19.
- IELTS. (2012). IELTS: Guide for teachers. Retrieved on 14-01-2017 from https://www.britishcouncil.it/sites/default/files/ielts_guide_for_teachers_italy.pdf
- International Phonetic Association. (2017). International phonetic alphabet chart. Retrieved on 29-12-2017 from <http://www.internationalphoneticassociation.org/content/ipa-chart>
- Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *Canadian Modern Language Review*, 64(4), 555-580.
- Isaacs, T., & Harding, L. (2017). Research timeline: Pronunciation assessment. *Language Teaching*, 50(3), 347-366.
- Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, 10(2), 135-159.
- Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics*, 32(1), 113-140.
- Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility. *Studies in Second Language Acquisition*, 34(03), 475-505.
- Isaacs, T., & Trofimovich, P. (2016). Second language pronunciation assessment: A look at the present and the future. In T. Isaacs, & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 259-271). Bristol: Multilingual Matters.
- Isaacs, T., Trofimovich, P., Yu, G., & Chereau, B. M. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS Research Reports Online Series*, 4, 1-48.
- Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 24-49.
- Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed.), *Conversation analysis: Studies from the first generation*. (pp. 13-31). Amsterdam: John Benjamins.

References

- Jenkins, J. (2000). *The phonology of English as an international language*. Oxford: Oxford University Press.
- Jenkins, J. (2014). *English as a lingua franca in the international university: The politics of academic English language policy*. Oxford: Routledge.
- Jesse, A., Poellmann, K., & Kong, Y. (2017). English listeners use suprasegmental cues to lexical stress early during spoken-word recognition. *Journal of Speech, Language, and Hearing Research, 60*, 190-198.
- Joe, J. N., Harmes, J. C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring: A mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy & Practice, 18*(3), 239-258.
- Johansson, S. (1978). *Studies of error gravity: Native reactions to errors produced by Swedish learners of English*. Volume 44 of Gothenburg Studies in English.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.
- Jones, N. (2000). Background to the validation of the ALTE can do project and the revised common European framework. *Research Notes, 2*, 11-13.
- Jun, H. G., & Li, J. (2010). Factors in raters' perceptions of comprehensibility and accentedness. In J. Levis, & K. LeVelle (Eds.), *Proceedings of the 1st pronunciation in second language learning and teaching conference* (pp. 53-66). Ames, IA: Iowa State University.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th Edition). Washington DC: American Council on Education/Praeger.
- Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly, 9*(3), 249-269.
- Kang, O. (2013a). Linguistic analysis of speaking features distinguishing general English exams at CEFR levels. *Research Notes, 52*, 40-48.
- Kang, O. (2013b). Relative impact of pronunciation features on ratings of non-native speakers' oral proficiency. In J. Levis, & K. LeVelle (Eds.), *Proceedings of the 4th pronunciation in second language learning and teaching conference, Aug. 2012*. (pp. 10-15) (pp. 10-15). Ames, IA: Iowa State University.
- Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal, 94*(4), 554-566.

- 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.
- Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, 12(3), 239-261.
- Kim, Y. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, 26(2), 187-217.
- Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. *Language Testing*, 28(2), 179-200.
- Knoch, U. (2016). What can pronunciation researchers learn from research into second language writing? In T. Isaacs, & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 54-71). Bristol: Multilingual Matters.
- Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, 118(2), 1038-1054.
- Kohler, K. (1982). Rhythmus im deutschen [Rhythm in German]. *Arbeitsberichte, Institut Für Phonetik Der Universität Kiel*, 19, 89-106.
- Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing*, 31(3), 329-348.
- Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Studies in language testing 9: Fairness and validation in language assessment: Selected papers from the 19th language testing research colloquium, Orlando, Florida* (pp. 1-14). Cambridge: Cambridge University Press.
- Ladd, D. R. (2008). *Intonational phonology*. Cambridge: Cambridge University Press.
- Laming, D. R. J. (1997). *The measurement of sensation*. Oxford: Oxford University Press.
- Lapadat, J. C. (2000). Problematizing transcription: Purpose, paradigm and quality. *International Journal of Social Research Methodology*, 3(3), 203-219.
- Lapadat, J. C., & Lindsay, A. C. (1999). Transcription in research and practice: From standardization of technique to interpretive positionings. *Qualitative Inquiry*, 5(1), 64-86.
- LaRossa, R. (2005). Grounded theory methods and qualitative family research. *Journal of Marriage and Family*, 67(4), 837-857.

References

- Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics*. Oxford: Oxford University Press.
- Laures, J. S., & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of Speech, Language, and Hearing Research*, 42(5), 1148-1156.
- Laver, J. (1994). *Principles of phonetics*. Cambridge: Cambridge University Press.
- Lee, Y., & Greene, J. (2007). The predictive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed Methods Research*, 1(4), 366-389.
- Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.
- Leighton, J., & Gierl, M. (2007). Verbal reports as data for cognitive diagnostic assessment. In J. Leighton, & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 146-172). Cambridge: Cambridge University Press.
- Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39(3), 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.
- Levis, J. M., & Wichmann, A. (2015). English intonation - form and meaning. In M. Reed, & J. M. Levis (Eds.), *The handbook of English pronunciation* (pp. 139-155). Oxford: Wiley Blackwell.
- Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, 28, 543-560.
- Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7(4), 328.
- Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, 16(2), 878.
- Linacre, J. M. (2013). A user's guide to FACETS Rasch-model computer programs. *Program Manual 3.67.0*. Retrieved on 10-12-2015 from: <https://www.winsteps.com/manuals.htm>
- Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, 3(2), 103-122.

- Linacre, J. M. (2012). *Many-facet rasch measurement: Facets tutorial*. Retrieved on 10-01-2016 from <https://www.winsteps.com/tutorials.htm>
- Linacre, J. M. (2014). *Facets (many-facet rasch measurement) Program, version 3.71.4*. Available from: <http://www.winsteps.com/facets.htm>
- Llurda, E. (2000). Effects of intelligibility and speaking rate on judgements of non-native speakers' personalities. *IRAL-International Review of Applied Linguistics in Language Teaching*, 38(3-4), 289-300.
- Luce, P., & McLennan, C. (2004). Spoken word recognition: The challenge of variation. In D. Pisoni, & R. Remez (Eds.), *The handbook of speech perception* (pp. 591-609). London: Wiley-Blackwell.
- Ludwig, J. (1982). Native-Speaker judgments of Second-Language learners' efforts at communication: A review. *The Modern Language Journal*, 66(3), 274-283.
- Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19(3), 246-276.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54-71.
- Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
- Lyle, J. (2003). Stimulated recall: A report on its use in naturalistic research. *British Educational Research Journal*, 29(6), 861-878.
- Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, , 173-190.
- Matthews, B., & Ross, L. (2010). *Research methods: A practical guide for the social sciences*. London: Pearson Education.
- May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, 26(3), 397-421.
- May, L. A. (2011). *Interaction in a paired speaking test*. Frankfurt am Main: Peter Lang.
- McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal*, 1(01).
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford: Blackwell.
- McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

References

- McNerney, M., & Mendelsohn, D. (1992). Suprasegmentals in the pronunciation class: Setting priorities. In P. Avery & S. Ehrlich (Eds.) *Teaching American English Pronunciation*, (pp. 185-196). Oxford: Oxford University Press
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan Publishing.
- Messick, S. (1992). Validity of test interpretation and use. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (6th edition). New York: Macmillan.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241-256.
- Mok, P., & Lee, S. I. (2008). Korean speech rhythm using rhythmic measures. *Proceedings of the 18th International Congress of Linguists (CIL18), Seoul, Korea*,
- Morgan, D. L. (2014). Pragmatism as a paradigm for social research. *Qualitative Inquiry*, 20(8), 1045-1053.
- Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, 23(2), 5-12.
- Munro, M. J., & Derwing, T. M. (1994). Evaluations of foreign accent in extemporaneous and read material. *Language Testing*, 11(3), 253-266.
- Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1), 73-97.
- Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, 38(3), 289-306.
- 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. (2015a). Intelligibility in research and practice: Teaching priorities. In M. Reed, & J. Levis (Eds.), *The handbook of English pronunciation* (pp. 377-396). Oxford: John Wiley & Sons.
- Munro, M. J., & Derwing, T. M. (2015b). A prospectus for pronunciation research in the 21st century: A point of view. *Journal of Second Language Pronunciation*, 1(1), 11-42.
- Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, 28(01), 111-131.

- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4), 386-422.
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part 2. *Journal of Applied Measurement*, 5(2), 189-227.
- Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, 24(3), 756.
- Nelson, C. L. (2008). Intelligibility since 1969. *World Englishes*, 27(3-4), 297-308.
- Nichols, J. (1988). Language study, international study, and education. *Profession*, 10-17.
- Nida, E. A. (1957). *Learning a foreign language*. Ann Arbor: Friendship.
- Nolan, F. (2006). Intonation. In Aarts, B., & McMahon, A. (Eds.), *The handbook of English linguistics* (pp. 433-458). Oxford: Blackwell.
- North, B. (2000). *The development of a common framework scale of language proficiency*. Bern: Peter Lang.
- North, B. (2005). The CEFR levels and descriptor scale. In Taylor, L. & Weir C. J. (Eds) *Studies in language testing 27: Multilingualism in assessment*. Cambridge: Cambridge University Press.
- North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, 15(2), 217-262.
- O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing*, 19(2), 169-192.
- O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, 19(3), 277-295.
- O'Connor, J. D. (1965). The perception of time intervals. *Progress Report*, 2, 11-15.
- Onwuegbuzie, A. J., & Combs, J. P. (2010). Emergent data analysis techniques in mixed methods research: A synthesis. In A. Tashakkori, & C. Teddlie (Eds.), *SAGE handbook of mixed methods in social and behavioral research* (2nd ed., pp. 397-430). Thousand Oaks, CA: Sage.
- Ordin, M., Polyanskaya, L., & Ulbrich, C. (2011). Acquisition of timing patterns in second language. *Interspeech*, 1129-1132.
- Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, 30(2), 143-154.

References

- Oxenden, C., & Latham-Koenig, C. (2013). *English file: Intermediate*. Oxford: Oxford University Press.
- Pan, M. (2015). *Nonverbal delivery in speaking assessment: From an argument to A rating scale formulation and validation*. London: Springer.
- Pearson. (2016a). Global scale of English assessment framework. Retrieved on 12-09-2017 from https://prodengcom.s3.amazonaws.com/GSE_Assessment_FINAL_May_2016.pdf#public
- Pearson. (2016b). Global scale of English learning objectives for adult learners. Retrieved on 12-09-2017 from https://prodengcom.s3.amazonaws.com/GSE_LO_AdultLearners_0415.pdf#public
- Pearson (2017a). *PTE Academic: Score guide, version 8*. Retrieved on 17-11-2017 from: <https://pearsonpte.com/wp-content/uploads/2017/08/Score-Guide.pdf>
- Pearson (2017b). Speaking: 1.3 repeat sentence. Accessed on 12-09-2017 from <https://pearsonpte.com/the-test/format/english-speaking-writing/repeat-sentence/>
- Pearson (2017c). Speaking: 1.4 describe image. Accessed on 12-09-2017 from <https://pearsonpte.com/the-test/format/english-speaking-writing/describe-image/>
- Pearson (2017d). Speaking: 1.6 answer short question. Accessed on 12-09-2017 from <https://pearsonpte.com/the-test/format/english-speaking-writing/answer-short-question/>
- Pearson (2017e). Why choose PTE academic. Accessed on 12-09-2017 from <https://pearsonpte.com/why-pte-academic/>
- Piccardo, E. (2016). *Common European framework of reference for languages: Learning, teaching, assessment: Phonological scale revision process report*. Strasbourg: Council of Europe.
- Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly*, 35(2), 233-255.
- Pike, K. L. (1945). *The intonation of American English*. (1st ed.). Ann-Arbor: University of Michigan Press.
- Pike, K. L. (1946). *The intonation of American English*. (2nd ed.). Ann-Arbor: University of Michigan Press.
- Pinget, A., Bosker, H. R., Quené, H., & de Jong, N. H. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing*, 31(3), 349-365.

- Pisoni, D. B. (1997). Some thoughts on “normalization” in speech perception. In K. Johnson, & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9-32). London: Academic Press.
- Pittam, J., & Scherer, K. R. (1993). *Vocal expression and communication of emotion*. New York: Guilford Press.
- Plano Clark, V. L., Garrett, A. L., & Leslie-Pelecky, D. L. (2010). Applying three strategies for integrating quantitative and qualitative databases in a mixed methods study of a nontraditional graduate education program. *Field Methods*, 22(2), 154-174.
- 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 research testing colloquium, cambridge and arnhem* (pp. 74-91). Cambridge: Cambridge University Press.
- Prieto, G., & Nieto, E. (2014). Analysis of rater severity on written expression exam using many faceted Rasch measurement. *Psicologica: International Journal of Methodology and Experimental Psychology*, 35(2), 385-397.
- QSR International. (2012). *NVivo qualitative data analysis software version* (10th ed.) Accessed from: <http://www.qsrinternational.com/nvivo/nvivo-products>
- R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rajadurai, J. (2007). Intelligibility studies: A consideration of empirical and ideological issues. *World Englishes*, 26(1), 87-98.
- Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *The Journal of the Acoustical Society of America*, 105(1), 512-521.
- Reemann, E., Alas, E., & Liiv, S. (2013). Interviewer behaviour during oral proficiency interviews: A gender perspective. *Eesti Rakenduslingvistika Ühingu Aastaraamat*, (9), 209-226.
- Richards, M. G. (2016). *Not all word stress errors are created equal: Validating an English word stress error gravity hierarchy*. PhD thesis: Iowa state university. Ann Arbor: MI: ProQuest.
- Roach, P. (2000). *English phonetics and phonology: A practical course* (3rd edition). 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(3), 395-412.

References

- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, , 696-735.
- Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, 37(2), 217-240.
- Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, 38(4), 439-462.
- Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (2016). Re-examining phonological and lexical correlates of second language comprehensibility: The role of rater experience. In T. Isaacs, & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 141-155). Bristol: Multilingual Matters.
- Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th language testing research colloquium, Orlando, Florida* (pp. 129-152). Cambridge: Cambridge University Press.
- Saldaña, J. (2015). *The coding manual for qualitative researchers*. London: Sage.
- Salzberger, T. (2015). The validity of polytomous items in the Rasch model: The role of statistical evidence of the threshold order. *Psychological Test and Assessment Modeling*, 3, 377-395.
- Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, 24(3), 355-390.
- Schairer, K. E. (1992). Native speaker reaction to non-native speech. *The Modern Language Journal*, 76(3), 309-319.
- Schmid, M. S., & Hopp, H. (2014). Comparing foreign accent in L1 attrition and L2 acquisition: Range and rater effects. *Language Testing*, 31(3), 367-388.
- Schneider, B. A., Daneman, M., & Pichora-Fuller, M. K. (2002). Listening in aging adults: From discourse comprehension to psychoacoustics. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale*, 56(3), 139.
- Scott, D. R., Isard, S. D., & de Boysson-Bardies, B. (1985). Perceptual isochrony in English and in French. *Journal of Phonetics*, 13(2), 155-162

- Seong, C. (1995). *The experimental phonetic study of standard current Korean speech rhythm: With respect to its temporal structure*. (Unpublished PhD Thesis). Seoul National University, Seoul: South Korea.
- Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing: Studies in language testing* 26. Cambridge: Cambridge University Press.
- Shoemaker, E., & Rast, R. (2013). Extracting words from the speech stream at first exposure. *Second Language Research*, 29(2), 165-183.
- Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, 11(2), 99-123.
- Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, 15, 188-211.
- Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, 76(1), 27-33.
- Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, 1(3), 185-211.
- Slowiaczek, L. M. (1990). Effects of lexical stress in auditory word recognition. *Language and Speech*, 33(1), 47-68.
- Small, L. H., Simon, S. D., & Goldberg, J. S. (1988). Lexical stress and lexical access: Homographs versus nonhomographs. *Attention, Perception, & Psychophysics*, 44(3), 272-280.
- Smith, L. E. (1992). Spread of English and issues of intelligibility. In B. B. Kachru (Ed.), *The other tongue: English across cultures* (2nd ed., pp. 75-90). Urbana, IL: University of Illinois Press.
- Smith, L. E., & Bisazza, J. A. (1982). The comprehensibility of three varieties of English for college students in seven countries. *Language Learning*, 32(2), 259-269.
- Smith, L. E., & Nelson, C. L. (1985). International intelligibility of English: Directions and resources. *World Englishes*, 4(3), 333-342.
- Smith, L. E., & Rafiqzad, K. (1979). English for cross-cultural communication: The question of intelligibility. *TESOL Quarterly*, 13(3), 371-380.
- Someren, M. v., Barnard, Y. F., & Sandberg, J. A. (1994). *The think aloud method: A practical approach to modelling cognitive processes*. London: Academic Press.

References

- Southwood, M. H., & Flege, J. E. (1999). Scaling foreign accent: Direct magnitude estimation versus interval scaling. *Clinical Linguistics & Phonetics*, 13(5), 335-349.
- Speer, S., & Blodgett, A. (2006). Prosody. In B. A. Speer S. (Ed.), *Handbook of psycholinguistics* (pp. 505-537). Amsterdam, The Netherlands: Elsevier.
- Stemler, S. E., & Tsai, J. (2008). Best practices in estimating interrater reliability. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29-49). Thousand Oaks, CA: Sage.
- Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, 53(1), 1-26.
- Swan, M., & Smith, B. (2001). *Learner English: A teacher's guide to interference and other problems* (2nd ed.). Cambridge: Cambridge University Press.
- Sweet, H. (1906). *A primer of phonetics* (3rd ed.). Oxford: Clarendon Press.
- Tajima, K., Port, R., & Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. *Journal of Phonetics*, 25(1), 1-24.
- Takala, S. (2010). Putting the CEFR to good use: Activities and outcomes in Finland. *IATEFL TEA SIG/EALTA Conference Proceedings, 2010*, Barcelona. 96-105.
- Taylor, L. (2006). The changing landscape of English: implications for English language assessment. *ELT Journal*, 60(1), 51-60.
- Taylor, L. (2011). Introduction. In L. Taylor (Ed.), *Studies in language testing 30: Examining speaking* (pp. 1-35). Cambridge: Cambridge University Press.
- Taylor, L., & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.), *Studies in language testing 30: Examining speaking* (pp. 171-233). Cambridge: Cambridge University Press.
- Taylor, S. J., Bogdan, R., & DeVault, M. L. (2016). *Introduction to qualitative research methods a guidebook and resource* (4th ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.
- Teddlie, C., & Yu, F. (2007). Mixed methods sampling a typology with examples. *Journal of Mixed Methods Research*, 1(1), 77-100.
- Tench, P. (1997). Towards a design of a pronunciation test. *Speak Out*, 20, 29-43.
- Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning*, 41(2), 177-204.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

- Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, 15(04), 905-916.
- Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, 49(1), 3-12.
- Urquhart, C., Lehmann, H., & Myers, M. D. (2010). Putting the 'theory' back into grounded theory: Guidelines for grounded theory studies in information systems. *Information Systems Journal*, 20(4), 357-381.
- Vaissière, J. (2005). Perception of intonation. In D. B. Pisoni, & R. E. Remez (Eds.), *Handbook of speech perception* (pp. 236-263). Oxford: Blackwell.
- Van Donselaar, W., Koster, M., & Cutler, A. (2005). Exploring the role of lexical stress in lexical recognition. *The Quarterly Journal of Experimental Psychology Section A*, 58(2), 251-273.
- Van Moere, A. (2013). Raters and ratings. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1358-1374). Hoboken, NJ: Wiley.
- Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly*, 12(3), 283-304.
- Weigle, S. C., Boldt, H., & Valsecchi, M. (2003). Effects of task and rater background on the evaluation of ESL student writing: A pilot study. *TESOL Quarterly*, 37(2), 345-354.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263-287.
- Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.
- Wells, J. C. (2006). *English intonation: An introduction*. Cambridge: Cambridge University Press.
- Wennerstrom, A. (1998). Intonation as cohesion in academic discourse. *Studies in Second Language Acquisition*, 20(01), 1-25.
- Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. Oxford: Oxford University Press.
- Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., & Mattys, S. L. (2010). How stable are acoustic metrics of contrastive speech rhythm? *The Journal of the Acoustical Society of America*, 127(3), 1559-1569.

References

- Wilkins, C. P., (1978). Proposals for level definitions. In Trim, J. L. M. *Some possible lines of development of an overall structure for a European unit/credit scheme for foreign language learning by adults*. Appendix C, Strasbourg: Council of Europe, 71-81
- Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly*, 47(4), 762-789.
- Winke, P., Gass, S., & Myford, C. (2011). The relationship between raters' prior language study and the evaluation of foreign language speech samples. *ETS Research Report Series*, 2
- Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2), 231-252.
- Wolfson, N. (1989). *Perspectives: Sociolinguistics and TESOL*. Boston MA: Heinle and Heinle.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.
- Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, 31(4), 501-527.
- Yates, L., Zielinski, B., & Pryor, E. (2011). The assessment of pronunciation and the new IELTS pronunciation scale. *IELTS Research Reports*, 12, 1-46.
- Zhang, Y., & Francis, A. (2010). The weighting of vowel quality in native and non-native listeners' perception of English lexical stress. *Journal of Phonetics*, 38(2), 260-271.
- Zhang, Y., Nissen, S. L., & Francis, A. L. (2008). Acoustic characteristics of English lexical stress produced by native Mandarin speakers. *The Journal of the Acoustical Society of America*, 123(6), 4498-4513.
- Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, 28(1), 31-50.
- Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the college English test-spoken English test (CET-SET). *Assessment in Education: Principles, Policy & Practice*, 21(3), 306-325.
- Zheng, Y. (2009). Protocol analysis in the validation of language tests: Potential of the method, state of the evidence. *International Journal of Pedagogies and Learning*, 5(1), 124-137.
- Zielinski, B. (2015). The segmental/suprasegmental debate. In M. Reed, & J. M. Levis (Eds.), *The handbook of English pronunciation* (pp. 397-412). Oxford: Wiley Blackwell.

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, 36(1), 69-84.

Appendices

Appendix A: Pronunciation Frameworks

CEFR Phonological Control Grid (2001)

	PHONOLOGICAL CONTROL
C2	As C1
C1	<i>Can vary intonation and place sentence stress correctly in order to express finer shades of meaning.</i>
B2	<i>Has acquired a clear, natural, pronunciation and intonation.</i>
B1	<i>Pronunciation is clearly intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur.</i>
A2	<i>Pronunciation is generally clear enough to be understood despite a noticeable foreign accent, but conversational partners will need to ask for repetition from time to time.</i>
A1	<i>Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by native speakers used to dealing with speakers of his/her language group.</i>

(Council of Europe, 2001: 117)

Suprasegmental Features Included in the 2017 CEFR

	OVERALL PHONOLOGICAL CONTROL	SOUND ARTICULATION	PROSODIC FEATURES
C2	Can employ the full range of phonological features in the target language with a high level of control – including prosodic features such as word and sentence stress, rhythm and intonation – so that the finer points of his/her message are clear and precise. Intelligibility and effective conveyance of and enhancement of meaning are not affected in any way by features of accent that may be retained from other language(s).	Can articulate virtually all the sounds of the target language with clarity and precision.	Can exploit prosodic features (e.g. stress, rhythm and intonation) appropriately and effectively in order to convey finer shades of meaning (e.g. to differentiate and emphasise).
C1	Can employ the full range of phonological features in the target language with sufficient control to ensure intelligibility throughout. Can articulate virtually all the sounds of the target language; some features of accent retained from other language(s) may be noticeable, but they do not affect intelligibility.	Can articulate virtually all of the sounds of the target language with a high degree of control. He/she can usually self-correct if he/she noticeably mispronounces a sound.	Can produce smooth, intelligible spoken discourse with only occasional lapses in control of stress, rhythm and/or intonation, which do not affect intelligibility or effectiveness. Can vary intonation and place stress correctly in order to express precisely what he/she means to say.
B2	Can generally use appropriate intonation, place stress correctly and articulate individual sounds clearly; accent tends to be influenced by other language(s) he/she speaks, but has little or no effect on intelligibility.	Can articulate a high proportion of the sounds in the target language clearly in extended stretches of production; is intelligible throughout, despite a few systematic mispronunciations. Can generalise from his/her repertoire to predict the phonological features of most unfamiliar words (e.g. word stress) with reasonable accuracy (e.g. whilst reading).	Can employ prosodic features (e.g. stress, intonation, rhythm) to support the message he/she intends to convey, though with some influence from other languages he/she speaks.
B1	Pronunciation is generally intelligible; can approximate intonation and stress at both utterance and word levels. However, accent is usually influenced by other language(s) he/she speaks.	Is generally intelligible throughout, despite regular mispronunciation of individual sounds and words he/she is less familiar with.	Can convey his/her message in an intelligible way in spite of a strong influence on stress, intonation and/or rhythm from other language(s) he/she speaks.

A2	<p>Pronunciation is generally clear enough to be understood, but conversational partners will need to ask for repetition from time to time. A strong influence from other language(s) he/she speaks on stress, rhythm and intonation may affect intelligibility, requiring collaboration from interlocutors. Nevertheless, pronunciation of familiar words is clear.</p>	<p>Pronunciation is generally intelligible when communicating in simple everyday situations, provided the interlocutor makes an effort to understand specific sounds. Systematic mispronunciation of phonemes does not hinder intelligibility, provided the interlocutor makes an effort to recognise and adjust to the influence of the speaker's language background on pronunciation.</p>	<p>Can use the prosodic features of everyday words and phrases intelligibly, in spite of a strong influence on stress, intonation and/or rhythm from other language(s) he/she speaks. Prosodic features (e.g. word stress) are adequate for familiar, everyday words and simple utterances.</p>
A1	<p>Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by interlocutors used to dealing with speakers of the language group concerned. Can reproduce correctly a limited range of sounds as well as the stress on simple, familiar words and phrases.</p>	<p>Can reproduce sounds in the target language if carefully guided. Can articulate a limited number of sounds, so that speech is only intelligible if the interlocutor provides support (e.g. by repeating correctly and by eliciting repetition of new sounds).</p>	<p>Can use the prosodic features of a limited repertoire of simple words and phrases intelligibly, in spite of a very strong influence on stress, rhythm, and/or intonation from other language(s) he/she speaks; his/her interlocutor needs to be collaborative.</p>

(Council of Europe, 2017: 134-5)

Horner's Proposed New Phonological Control Grid

Table 2: Proposed new phonological control grid descriptors	
C2	<ul style="list-style-type: none"> • Speaker is easily understood. • Mispronunciations are rare. • Sentence stress is used successfully most of the time. • Intonation is used successfully most of the time.
C1	<ul style="list-style-type: none"> • Speaker is easily understood. • Mispronunciations are rare. • Sentence stress is used successfully most of the time. • Intonation is used but not always effectively.
B2	<ul style="list-style-type: none"> • Speaker is understood. • Mispronunciations occur but do not interfere with understanding. • Sentence stress is used but not always successfully. • Basic intonation patterns are used, but not always successfully.
B1	<ul style="list-style-type: none"> • Sufficient control of sounds to be understandable. • Sufficient control of word stress to be understandable • Mispronunciations occur, but only occasionally interfere with understanding.
A2	<ul style="list-style-type: none"> • Sufficient command of sounds and word stress to be understandable, but with some difficulty. • The interlocutor may need to ask for repetition or clarification.
A1	<ul style="list-style-type: none"> • Sufficient command of sounds to be understandable, but not all of the time and with some difficulty. • Sufficient command of word stress to be understandable, but not all of the time and with some difficulty. • The interlocutor will need to ask for repetition or clarification.

(Horner, 2010: 56)

Appendix B: Speaking Rating Scale Descriptors

IELTS: Speaking Band Descriptor for Pronunciation (Public Version)

Band	Pronunciation
9	<ul style="list-style-type: none"> • Uses a full range of pronunciation features with precision and subtlety • Sustains flexible use of features throughout • Is effortless to understand
8	<ul style="list-style-type: none"> • Uses a wide range of pronunciation features • Sustains flexible use of features, with only occasional lapses • Is easy to understand throughout; 1 accent has minimal effect on intelligibility
7	<ul style="list-style-type: none"> • Shows all the positive features of band 6 and some, but not all, of the positive features of band 8
6	<ul style="list-style-type: none"> • Uses a range of pronunciation features with mixed control • Shows some effective use of features but this is not sustained • Can generally be understood throughout, though mispronunciation of individual words or sounds reduces clarity at times
5	<ul style="list-style-type: none"> • Shows all the positive features of band 4 and some, but not all, of the positive features of band 6
4	<ul style="list-style-type: none"> • 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
3	<ul style="list-style-type: none"> • Shows some of the features of band 2 and some, but not all, of the positive features of band 4
2	<ul style="list-style-type: none"> • Speech is often unintelligible
1	
0	

(IELTS, 2012)

Cambridge English: Overall Speaking Scale (Cambridge English, 2016b: 83)

	Grammatical Resource	Lexical Resource	Discourse Management	Pronunciation	Interactive Communication
	<ul style="list-style-type: none"> Maintains control of a wide range of grammatical forms and uses them with flexibility. 	<ul style="list-style-type: none"> Uses a wide range of appropriate vocabulary with flexibility to give and exchange views on unfamiliar and abstract topics. 	<ul style="list-style-type: none"> Produces extended stretches of language with flexibility and ease and very little hesitation. Contributions are relevant, coherent, varied and detailed. Makes full and effective use of a wide range of cohesive devices and discourse markers. 	<ul style="list-style-type: none"> Is intelligible. Phonological features are used effectively to convey and enhance meaning. 	<ul style="list-style-type: none"> Interacts with ease by skilfully interweaving his/her contributions into the conversation. Widens the scope of the interaction and develops it fully and effectively towards a negotiated outcome.
C2	<ul style="list-style-type: none"> Maintains control of a wide range of grammatical forms. 	<ul style="list-style-type: none"> Uses a wide range of appropriate vocabulary to give and exchange views on unfamiliar and abstract topics. 	<ul style="list-style-type: none"> Produces extended stretches of language with ease and with very little hesitation. Contributions are relevant, coherent and varied. Uses a wide range of cohesive devices and discourse markers. 	<ul style="list-style-type: none"> Is intelligible. Intonation is appropriate. Sentence and word stress is accurately placed. Individual sounds are articulated clearly. 	<ul style="list-style-type: none"> Interacts with ease, linking contributions to those of other speakers. Widens the scope of the interaction and negotiates towards an outcome.
C1	<ul style="list-style-type: none"> Shows a good degree of control of a range of simple and some complex grammatical forms. 	<ul style="list-style-type: none"> Uses a range of appropriate vocabulary to give and exchange views on familiar and unfamiliar topics. 	<ul style="list-style-type: none"> Produces extended stretches of language with very little hesitation. Contributions are relevant and there is a clear organisation of ideas. Uses a range of cohesive devices and discourse markers. 	<ul style="list-style-type: none"> Is intelligible. Intonation is appropriate. Sentence and word stress is accurately placed. Individual sounds are articulated clearly. 	<ul style="list-style-type: none"> Initiates and responds appropriately, linking contributions to those of other speakers. Maintains and develops the interaction and negotiates towards an outcome.
Grammar and Vocabulary					
B2	<ul style="list-style-type: none"> Shows a good degree of control of simple grammatical forms, and attempts some complex grammatical forms. Uses appropriate vocabulary to give and exchange views, on a range of familiar topics. 		<ul style="list-style-type: none"> Produces extended stretches of language despite some hesitation. Contributions are relevant and there is very little repetition. Uses a range of cohesive devices. 	<ul style="list-style-type: none"> Is intelligible. Intonation is generally appropriate. Sentence and word stress is generally accurately placed. Individual sounds are generally articulated clearly. 	<ul style="list-style-type: none"> Initiates and responds appropriately. Maintains and develops the interaction and negotiates towards an outcome with very little support.
B1	<ul style="list-style-type: none"> Shows a good degree of control of simple grammatical forms. Uses a range of appropriate vocabulary when talking about familiar topics. 		<ul style="list-style-type: none"> Produces responses which are extended beyond short phrases, despite hesitation. Contributions are mostly relevant, but there may be some repetition. Uses basic cohesive devices. 	<ul style="list-style-type: none"> Is mostly intelligible, and has some control of phonological features at both utterance and word levels. 	<ul style="list-style-type: none"> Initiates and responds appropriately. Keeps the interaction going with very little prompting and support.
A2	<ul style="list-style-type: none"> Shows sufficient control of simple grammatical forms. Uses appropriate vocabulary to talk about everyday situations. 			<ul style="list-style-type: none"> Is mostly intelligible, despite limited control of phonological features. 	<ul style="list-style-type: none"> Maintains simple exchanges, despite some difficulty. Requires prompting and support.
A1	<ul style="list-style-type: none"> Shows only limited control of a few grammatical forms. Uses a vocabulary of isolated words and phrases. 			<ul style="list-style-type: none"> Has very limited control of phonological features and is often unintelligible. 	<ul style="list-style-type: none"> Has considerable difficulty maintaining simple exchanges. Requires additional prompting and support.

TOEFL iBT

Independent Speaking Scale

SCORE	GENERAL DESCRIPTION	DELIVERY	LANGUAGE USE	TOPIC DEVELOPMENT
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Generally well-paced flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility.	The response demonstrates effective use of grammar and vocabulary. It exhibits a fairly high degree of automaticity with good control of basic and complex structures (as appropriate). Some minor (or systematic) errors are noticeable but do not obscure meaning.	Response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear (or clear progression of ideas).
3	The response addresses the task appropriately but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected).	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. This may affect overall fluency, but it does not seriously interfere with the communication of the message.	Response is mostly coherent and sustained and conveys relevant ideas/information. Overall development is somewhat limited, usually lacks elaboration or specificity. Relationships between ideas may at times not be immediately clear.
2	The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/or overall coherence occur; meaning may be obscured in places. A response at this level is characterized by at least two of the following:	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/pace; meaning may be obscured in places.	The response demonstrates limited range and control of grammar and vocabulary. These limitations often prevent full expression of ideas. For the most part, only basic sentence structures are used successfully and spoken with fluidity. Structures and vocabulary may express mainly simple (short) and/or general propositions, with simple or unclear connections made among them (serial listing, conjunction, juxtaposition).	The response is connected to the task, though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed with limited elaboration (details and support). At times relevant substance may be vaguely expressed or repetitious. Connections of ideas may be unclear.
1	The response is very limited in content and/or coherence or is only minimally connected to the task, or speech is largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation, stress and intonation difficulties cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limit or prevent expression of ideas and connections among ideas. Some low-level responses may rely heavily on practiced or formulaic expressions.	Limited relevant content is expressed. The response generally lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete the task and may rely heavily on repetition of the prompt.
0	Speaker makes no attempt to respond OR response is unrelated to the topic.			

(ETS, 2014)

Integrated Speaking Scale

SCORE	GENERAL DESCRIPTION	DELIVERY	LANGUAGE USE	TOPIC DEVELOPMENT
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Speech is generally clear, fluid, and sustained. It may include minor lapses or minor difficulties with pronunciation or intonation. Pace may vary at times as the speaker attempts to recall information. Overall intelligibility remains high.	The response demonstrates good control of basic and complex grammatical structures that allow for coherent, efficient (automatic) expression of relevant ideas. Contains generally effective word choice. Though some minor (or systematic) errors or imprecise use may be noticeable, they do not require listener effort (or obscure meaning).	The response presents a clear progression of ideas and conveys the relevant information required by the task. It includes appropriate detail, though it may have minor errors or minor omissions.
3	The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, but it exhibits minor difficulties with pronunciation, intonation, or pacing and may require some listener effort at times. Overall intelligibility remains good, however.	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. Such limitations do not seriously interfere with the communication of the message.	The response is sustained and conveys relevant information required by the task. However, it exhibits some incompleteness, inaccuracy, lack of specificity with respect to content, or choppiness in the progression of ideas.
2	The response is connected to the task, though it may be missing some relevant information or contain inaccuracies. It contains some intelligible speech, but at times problems with intelligibility and/or overall coherence may obscure meaning. A response at this level is characterized by at least two of the following:	Speech is clear at times, though it exhibits problems with pronunciation, intonation, or pacing and so may require significant listener effort. Speech may not be sustained at a consistent level throughout. Problems with intelligibility may obscure meaning in places (but not throughout).	The response is limited in the range and control of vocabulary and grammar demonstrated (some complex structures may be used, but typically contain errors). This results in limited or vague expression of relevant ideas and imprecise or inaccurate connections. Automaticity of expression may only be evident at the phrasal level.	The response conveys some relevant information but is clearly incomplete or inaccurate. It is incomplete if it omits key ideas, makes vague reference to key ideas, or demonstrates limited development of important information. An inaccurate response demonstrates misunderstanding of key ideas from the stimulus. Typically, ideas expressed may not be well connected or cohesive so that familiarity with the stimulus is necessary to follow what is being discussed.
1	The response is very limited in content or coherence or is only minimally connected to the task. Speech may be largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation and intonation problems cause considerable listener effort and frequently obscure meaning. Delivery is choppy, fragmented, or telegraphic. Speech contains frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limit (or prevent) expression of ideas and connections among ideas. Some very low-level responses may rely on isolated words or short utterances to communicate ideas.	The response fails to provide much relevant content. Ideas that are expressed are often inaccurate, limited to vague utterances, or repetitions (including repetition of prompt).
0	Speaker makes no attempt to respond OR response is unrelated to the topic.			

(ETS, 2014)

Pearson Test of English: Academic, Scoring Criteria: Pronunciation and Oral Fluency

Pronunciation	
5 Native-like	All vowels and consonants are produced in a manner that is easily understood by regular speakers of the language. The speaker uses assimilation and deletions appropriate to continuous speech. Stress is placed correctly in all words and sentence-level stress is fully appropriate
4 Advanced	Vowels and consonants are pronounced clearly and unambiguously. A few minor consonant, vowel or stress distortions do not affect intelligibility. All words are easily understandable. A few consonants or consonant sequences may be distorted. Stress is placed correctly on all common words, and sentence level stress is reasonable
3 Good	Most vowels and consonants are pronounced correctly. Some consistent errors might make a few words unclear. A few consonants in certain contexts may be regularly distorted, omitted or mispronounced. Stress-dependent vowel reduction may occur on a few words
2 Intermediate	Some consonants and vowels are consistently mispronounced in a non-native like manner. At least 2/3 of speech is intelligible, but listeners might need to adjust to the accent. Some consonants are regularly omitted, and consonant sequences may be simplified. Stress may be placed incorrectly on some words or be unclear
1 Intrusive	Many consonants and vowels are mispronounced, resulting in a strong intrusive foreign accent. Listeners may have difficulty understanding about 1/3 of the words. Many consonants may be distorted or omitted. Consonant sequences may be non-English. Stress is placed in a non-English manner; unstressed words may be reduced or omitted and a few syllables added or missed
0 Non-English	Pronunciation seems completely characteristic of another language. Many consonants and vowels are mispronounced, misordered or omitted. Listeners may find more than 1/2 of the speech unintelligible. Stressed and unstressed syllables are realized in a non-English manner. Several words may have the wrong number of syllables
Oral fluency	
5 Native-like	Speech shows smooth rhythm and phrasing. There are no hesitations, repetitions, false starts or non-native phonological simplifications
4 Advanced	Speech has an acceptable rhythm with appropriate phrasing and word emphasis. There is no more than one hesitation, one repetition or a false start. There are no significant non-native phonological simplifications
3 Good	Speech is at an acceptable speed but may be uneven. There may be more than one hesitation, but most words are spoken in continuous phrases. There are few repetitions or false starts. There are no long pauses and speech does not sound staccato
2 Intermediate	Speech may be uneven or staccato. Speech (if ≥ 6 words) has at least one smooth three-word run, and no more than two or three hesitations, repetitions or false starts. There may be one long pause, but not two or more

1 Limited	Speech has irregular phrasing or sentence rhythm. Poor phrasing, staccato or syllabic timing, and/or multiple hesitations, repetitions, and/or false starts make spoken performance notably uneven or discontinuous. Long utterances may have one or two long pauses and inappropriate sentence-level word emphasis
0 Disfluent	Speech is slow and labored with little discernable phrase grouping, multiple hesitations, pauses, false starts, and/or major phonological simplifications. Most words are isolated, and there may be more than one long pause

(Pearson, 2017a: 24)

Appendix C: Rating Sheet

Audio: [audio number]

Context: [e.g. 'the speaker is describing a picture of a food market']

How easy is it to understand the speaker?

1	2	3	4	5	6	7	8	9
Very								Very
difficult								easy

Notes

How well does the speech flow?

1	2	3	4	5	6	7	8	9
Very								Very
badly								well

Notes

How easy is it to recognise the words and phrases?

1	2	3	4	5	6	7	8	9
Very								Very
difficult								easy

Notes

Appendix D: Sample Interview Transcriptions

This appendix contains transcriptions from one experienced rater, ER_02, and one non-experienced rater, NR_04, in response to a selection of test takers.

Participant: ER_02

Date of interview: 04/07/2016

Speaker 01:

ER_02: I went for a sort of middle mark there was a bit of strain on me I did need to concentrate quite a bit . if it's putting a lot of strain on me that's going to you know push it down to a much lower mark and at the top end if I'm almost . it's just like speaking to a native speaker you're not having to think about it you're just thinking about the meaning a lot more then I'd go for the top score . so I just went for the middle because there were times when I really had to . almost you know you're filling in the missing final consonants and you're really conscious of what he's saying and you're having to think about the actual individual sounds a lot more than in some of the other recordings

Researcher: anything else

ER_02: probably it feels quite disjointed as well quite sort of fragmented . you're almost waiting for the next word I don't know it's like a young child reading . you know where they're putting their finger on each word so sometimes it's hard to get a global . you know what is the sentence all about because you're focusing on the individual sounds a lot more than you would normally . sort of thing . so I think it probably distracts you because your focus is almost in the wrong place as a listener . you should be listening for the sort of overall message and meaning but you're

still having to focus on the individual sounds which makes it a much slower process to actually understand what they're saying

Researcher: the focus is in the wrong place?

ER_02: yea it's the lack of intonation I think as well that makes it harder for you to get the message because it's fairly flat sort of staccato . I wrote down the word staccato at one point because it's this kind of der der der der . so you kind of . you find it hard to see chunks . you know . you've still very much got that individual word focus I think . I think you're almost having to piece it back together for him . so that makes it so that harder work as a listener I think . I think if there was some attempt at intonation it would make it a lot easier because you'd be getting those emphasised words . obviously the semantic words which erm would make it easier to follow . as it is I mean it's not a massive strain it's not like I'm not understanding any of it but I am having to make an effort to get the message

Researcher: anything else about that speaker

ER_02: apart from individual sounds which you might want to focus down on . as a teacher you'd want to focus on but as an assessor you become obviously . you know experience teaches you what to expect from different language speaking groups doesn't it . so you know that certain nationalities are going to have a problem with certain words erm . and I think . I don't know whether that helps you or not . whether you kind of think erm . I can't really mark them down because they can't pronounce /l/ and /r/ properly or /s/ and /f/ . do you know what I mean

Speaker 03

ER_02: I think generally she communicated very well and you know it had a good flow to it . there were one or two individual sounds . you know . the think /θɪnk/ fink /fɪnk/ . but I think native speakers . because a lot of native speakers get that wrong as well

don't they . you know . I think so . so I think people are used to making allowances for that one . I think the major issue would probably be with the intonation pattern . I've just put a kind of unnatural intonation pattern for native speaker listeners because the French . the kind of the rhythm of the language is so different from English isn't it . so I think it would be more a problem with the intonation pattern than the actual the individual words and utterances in this one but I think she conveyed things fairly clearly I think in this case . I was a bit more kind of . I didn't have to listen and try hard . quite so hard as with the other ones

Researcher: anything else on the pronunciation?

ER_02: it's the sort of der der der der der der der der der ((imitating the stress)). isn't it . I mean obviously French and English are so different in terms of intonation pattern that it's hard for any French speaker to mimic English intonation isn't it . and as a listener you could . you're picking out the wrong words aren't you . you're getting the wrong things thrown at you so it can be quite confusing . so there's no attempt at emphasising really important words so it does make it difficult yea

Speaker 04:

ER_02: I think in this case she used so many native speaker type discourse markers and connective . and the 'wells' and the 'I means' . just made it so much easier to follow what she was saying . it kind of took a lot of the strain off listening although there were still issues with some individual sounds . just that natural breaking it up explaining going back erm . it just made it a lot easier to follow and the message came across really really clearly . that's why I gave it an eight . it was just the naturalness of it . if naturalness is a word

Researcher: what do you remember making it natural?

ER_02: it's the little filling words isn't it the you know before when they hesitated they just hesitated . there was the gap and you might think . oh have they finished or what's coming next . but when you've got the 'I means' and you know it just it keeps your attention and it links it still keeps it linked together . which makes it easier to get the message much more clearly and I think when we're kind of examining at higher levels like CAE that's the kind of stuff that you're listening for . that native speaker type discourse where you can imagine having a conversation with them in a café or down the pub or something you know . it's just . it feels easy it feels comfortable it feels natural erm . I mean yea . you can take away big chunks of meaning from what she was saying just by the way she sort of broke it up . the way she delivered it . I might even go for a nine . there was a little bit of hesitancy but she always managed to fill it in with an appropriate connective or little discourse marker of some sort to keep it all together . so when you get those gaps and pauses and hesitations it really makes you feel they're struggling . they don't know the words they don't know the language . whereas when they're filling it in that natural way it just makes them flow it makes them seem like they're more in control doesn't it

Speaker 05:

ER_02: I found I was straining a little bit at times

Researcher: what was causing that strain

ER_02: probably the sort of stopping and starting . the hesitancy was interfering a bit . some individual sounds yea you expect that I suppose . but yea it caused me a bit of strain . not a great deal . I wrote some strain initially but I think it's more than some . I think it's making me really have to listen . fairly well connected she was using basic connectives but it was connected together and er . she was stopping but not hesitating . not stopping completely so yea she's trying . trying to connect

it together but maybe just needs to fill the gaps a bit more to make it feel a bit more comfortable less strain

Researcher: do you remember how she connected it?

ER_02: yea . she was emphasising some of the important words . so I got to the supermarket ((with exaggerated lexical stress on the first syllable)) . yea . sort of limited attempts at stressing words that were significant and again you can see that she's developing those skills . she's not quite there yet but she's going in the right direction yea

Researcher: okay . anything else on her?

ER_02: no . I think she communicated her main ideas pretty well . no . nothing major that would interfere . I just put some strain really not a great deal . not where I'm kind of . what was that

Speaker 09:

ER_02: okay . I was struggling more that time than I was the first time around . considerable strain at times with that one . individual sound level and utterance level I think . yea I was struggling at times . and again it was that flat intonation that really made it hard . everything was down der der der der der ((imitating falling pitch on each syllable)) . sort of staccato again . that sort of firing at you . yea but not giving you any clues as to the attitude or meaning really . yea it makes it hard . you're kind of doubly interpreting . at individual sound level you've got a problems and then you're trying to make sense of what they're actually trying to say at utterance level

Researcher: do you remember anything else?

ER_02: trying to think of it as a non-English teacher that would make it much harder . and there's also an attitude thing isn't there . we get so much about a person's attitude from their intonation in spoken English . it's very easy to get completely the wrong

idea about their attitude to what they're saying as well . they can sound bored fed up or even a bit angry with the wrong intonation which . in interaction can be quite difficult for them

Speaker 10:

ER_02: yea it's . erm . yea that naturalness again I think made it very easy to listen to and then you can see and then she was emphasising . you get a very /positive feeling and if you /win ber ber ber ((imitating the focal stress)) . so it was very easy to follow there . I mean grammatically not perfect . you can forgive all that because it was delivered in a very easy way . I didn't feel any strain at all there

Researcher: so this balance between grammar and natralness

ER_02: erm . I mean the German intonation it can be misinterpreted I think . it can sound . not angry . I mean people have described it as angry or rude . but I think she manages to try to overcome that a little bit by her little phrases that she throws in to kind of . you know you can see I think and well so she's really trying to break it up for us and to mark where she's adding information explaining or justifying . she's doing all those things to help us . yea I think that helps . helps a lot

Researcher: anything else on her

ER_02: no . I think the only . the only problem there is this sort of the wrong attitude that might come across . if you're not familiar with German intonation that might cause native speakers confusion . I sometimes feel they have to try extra hard to make it sound as if they're not angry not being rude . but yea I mean it's a small thing . I think she's trying really hard to interpret it for us to make it easy for us so follow what she's saying

Speaker 11:

ER_02: it's the intrusive sounds that make this hard isn't it . the young er . child er . lake er ((adding a schwa to the end of each of these words)) . those kinds of things which if you're not used to it . it can make it hard . I mean I was obviously very familiar with Italian so it's easy for me . I can see how that intrusive . you know it's the intrusive schwa on the end of the almost every single word . is it young or younger . lake or laker . that could be a problem for the non-linguist I guess . quite a bit of hesitancy and fragmentation as well wasn't there which again would make it hard for people to follow . slow deliberate erm . not necessarily a bad thing if she'd gabbled I think it would have been worse for people to try and follow erm . so yea I would say considerable strain for a non-linguist with that one . the intonation was . there's a beautiful /mountain ((imitating focal stress)) she was emphasising some of the key words but again I'm not sure it helped

Researcher: you said fragmented

ER_02: yea, I think for an unfamiliar listener the rhythm would be difficult to follow . it's interfering a little bit too much . it's almost the opposite problem with the clipped ends of words that you get with the sort of Cantonese Chinese speakers where they clip the ends of the words . you're actually emphasising the ends of the words and you're adding an extra kind of weak schwa sound as well so it's kind of an opposite extreme . erm I don't know . I don't know which is easier or harder whether it's the yer /jə/ or the younger /'jʌŋgə/ . what's easier to understand . I don't know . if you're not used to it I guess it's hard . I am very used to Italian so it's hard for me to be objective . but I think putting my mum's hat on she would struggle with that . I would say considerable strain and mostly because of fragmented delivery and the err on the end of everything

Researcher: anything else on her

ER_02: I'm not sure about the slowness . I mean it was reasonably fluent even though it was slow . yea it was reasonable well connected and fluent but just quite slow and deliberate . she was trying hard to make it clear wasn't she . I think that was the point but almost trying too hard I think probably

Speaker 12:

ER_02: yea there was some attempt to emphasise the key words book flowers she was cooking . yea it was kind of regular there was an attempt to stress the main words . fairly fragmented that would put a bit of strain

Researcher: kind of regular?

ER_02: yea on occasion . I think in some sentences that did come through didn't it . books and flowers and she was cooking . there was kind of an attempt to use intonation wasn't there . some individual sounds were still problematic . coffee coppee /kɔpi:/ . that kind of thing . again if you're not familiar with Japanese that kind of thing would make it harder . I think the main message . the utterances came across pretty clearly I think . I mean it was . yea it was fairly controlled in the sense that he managed to get those key words emphasises where he could

Participant: NR_04

Date of interview: 03/07/2016

Speaker 01:

NR_04: I think he sounds nervous and he really sounds like he's struggling . erm . I think he's probably another native Chinese speaker . and he sounds as though he's used to speaking quite fast but he's now speaking a language which isn't familiar to him . and he's trying to keep up that pace because that's what he's used to . but actually he's really struggling with it . and I think as a listener you pick up on that and you're really concentrating to get it . so in really concentrating I do know what he's saying and I can work out those words . and maybe familiarity with that accent helps but I think you would get it from the context . but you would have to really concentrate and it would help if you were looking at him as well . because his pronunciation is really quite poor . he's talking about the environment and he's talking about the city picture the urban picture looking very crowded . but he's not saying crowded . and he's not saying environment ((laughter)) that's not what that sound is coming out as erm . yea I just feel sorry for him because he sounds really bad . I think he's finding it really hard

Researcher: you said he sounds nervous

NR_04: erm . he's keeping up the speed but it doesn't . it's not like with some of the European speakers where that speed creates the flow . it's almost . there's almost a kind of a like a sort of staccato little breaks in that speed erm . it's sort of quite jittery . you could even think that he might be speeding up because of nervousness whereas if he was speaking more slowly the clarity would really be improved of what he is trying to say . it's as though he is speaking faster than the amount of thinking time he needs for the level of English that he's got

Researcher: anything more on him?

NR_04: no I just wrote that although I had to concentrate more I could recognise the words but you wouldn't be able to listen to him and do anything else at the same time because you have to concentrate on just what he's saying

Speaker 02:

NR_04: I found that one really difficult . I think the . his accent's really strong . and some of those words . I found the sounds coming from his native accent making some of those words really difficult

Researcher: can you say any more about that?

NR_04: I think there is that sort of quite an almost abrupt sound to the end of the sentence coming up at the last sound of the sentence . so there's almost like a questioning like a flow to it as he finishes it . it's almost as if he's asking for sort of like approval that he said it right it comes up at the end . but yea I don't know if that's coming from his native accent or if that's coming from a lack of confidence in what he's saying but there is like a pattern to how he says each sentence where there's a sort of . he's not saying umm but at the beginning he's drawing it out as like he's forming his thoughts and there's a speed up and a kind of flick up in the tone or the pitch at the end to make it sound almost like a question . and then he sort of each sentence as he says it follows that pattern like being drawn out and then going up

Researcher: and what was the impact on you?

NR_04: I think it makes it easier to identify sentence by sentence . but I think it probably sort of compounds that issue that he's pronouncing things in a way that's difficult to understand . and it's erm . it's not a sort of lilt that's helpful like that Spanish lady where it almost makes it more clear . his is more like you're kind of you're waiting

on that beginning of the sentence and then you're like oh yea I get it as he comes up at the end and it's almost like the understanding follows the pitch of how he's talking

Speaker 05:

NR_04: I found that one quite easy. I don't think the pronunciation is particularly good . it's very heavily influenced by her own language but . I think maybe it's familiar . the accent is familiar to me so that made it easier . also the increased speed made it easier . it joined things up where she has breaks and says umm it's more between a sentence than in the middle of a sentence . so that makes it easier . each sentence seems joined to each other umm because it's quite fast I sort of noted that even though she misses out words . it's still quite easy to follow . and the words that she seems to be missing were more sort of grammatical parts of speech . she doesn't say 'of' and things like that . rather than the kind of main building blocks of a sentence . the key vocabulary's there and that lets you know what she's talking about

Speaker 08:

NR_04: there were only one or two words that I couldn't get but again some things in isolation I wouldn't have been quite sure . it took the context to kind of maybe decide what sound he was aiming at because when the sounds starts the word isn't quite right and it's recognisable . when you know how it fits in with the sentence . so it's probably eastern European . it sounds erm the sounds from his language effecting how he's speaking again . it's obviously his level of language is lower and

I think there were words that he sort of . he's trying to recall and he sort of says one thing and then corrects himself or adds a bit of detail . when he says red chair then he calls it an armchair because he remembers what kind of chair that is . erm . so it . it required concentration to listen to erm sort of a little bit broken . he was trying to like recall other words erm . to say what he means and to give further detail to it . erm I suppose it almost feels slightly frustrating to listen to because you sort of feel like it's quite slow and it wants to kind of go somewhere

Speaker 09:

NR_04: I found that one quite difficult . because there was a lot of umms in it . which is something I do when I'm nervous . I think she sounds like she's probably nervous in the situation but also she's trying to . she's trying to describe these photos and she's trying to compare them to each other and I think she's trying to do a lot of things in what she's saying . and it almost . as she started better and she's almost getting more caught up in feeling like it's not going well because her umms get more frequent and it gets more difficult to follow . I don't really know what she was saying . she's not a very confident speaker . each word that she says is clear and understandable but because there's sort of three or four umms in between it's like breaking it

Speaker 10:

NR_04: it was really clear . I didn't think her accent . I wouldn't know where she was from . I would perhaps know that she wasn't a native speaker but her voice is very clear and it's not really influenced by her native sounds . so that doesn't detract from the

clarity of what she's saying in English . she's also obviously got a higher level of English than many of the others . she's got more vocabulary to draw on than to form her sentences and she's able to not just describe but I think at the end she also then starts putting in an opinion about it being better to play on a team than to play individually so . she's got enough language to talk around a topic more than just describe what she can see

Researcher: can you remember what it was that constituted clarity?

NR_04: yea I think a lot of the clarity is from the sounds being really well formed and there not being much sort of mispronunciation . I suppose the pronunciation is quite natural . it's quite . it feels almost like a native speaker and then there's also a confidence in being able to talk about the topic which adds to it as well

Researcher: confidence?

NR_04: quite a consistent speed . there aren't breaks in sentences while she thinks about what to say . it's . she just keeps talking . the voice doesn't go up and down and create those . sort of sentences which belong to each other in quite the same way . it's more . hers is more subtle . but I think the pace is consistent enough that it is quite clear that a sentence belongs to each other and that she knows where she's going with what she's saying . like her answer to the question as a whole seems quite confidently put together . when she talks about one she talks about the other and then she forms an opinion so there's a overall . like this makes sense as a paragraph of speech almost

Speaker 11:

NR_04: I found her really easy to understand and it's not necessarily because she's better than some of the others but there's something about her voice that's quite lilting . I

always feel like I can imagine this person . I feel like she gesticulates quite a lot while she's talking . erm and . yea there's something about the pace and just . about how she sort of elongated some of her vowel sounds maybe that comes from her language . I think she's probably a Spanish speaker . and yea just the way that made it feel like there was some proper cadence to the sentences and yea it was joined up and that made it much easier to recognise even though there were pronunciation and grammatical errors in it . it was still really easy to know what she was talking about

Researcher: you said lilting and proper cadence

NR_04: I think there's sort of almost like you would in a song . there was almost a quality of . and probably with the sentence structure of just there being the right kind of stresses and things to tie a sentence ((signalling up and down with hand)) together umm yea the gaps between words were sort of shortened and within the words seemed elongated because of the way she pronounced her vowels erm . it's quite slow . but because it felt joined up it wasn't frustratingly so to listen to . it felt like it was going somewhere and that it joined it up and that was it felt like a complete sentence . it is about the things are tied together so that the context of the sentence is so sort of built into it that it's not like some of the other ones where . I sort of felt like I know the word's recognisable and it's not quite right and if I listen to the next few words I'll know which it was meant to be . and then it'll make sense . it was as though that process didn't make sense for me because that word felt like it was part of the rest of the sentence so it was as though the context was almost already there before she said it . it sounds like a really abstract thing to say . it's there's something about tone and pitch there's something about maybe how her voice goes up and down to create that sentence which feels like it goes together so that when she's saying something and it's not quite right . but because you feel like you

recognise where that sentence is going like the understanding you already identify the sound before she's got to the next thing that she's got to say

Speaker 12:

NR_04: there were quite a few words where the pronunciation isn't quite right and without the context I'm not sure I would know what that word was . erm if it didn't tell me he was giving gifts I might have though he was giving her a pheasant rather than a present so . er . it doesn't make it feel . erm it makes it feel like there is a sentence but it is . yea it's as though he's thinking all the time as he's like planning each word . yea I know what he's saying but the words become quite individual because there's obviously that thought process going on the whole time . and if they weren't joined up with the other things I wouldn't have got all of them. . so I think it is a sentence but the broken-up-ness detracted from how well it came across

Appendix E: Preliminary Audio Rationales

Speaker 01

Chinese | FCE

Summary:

Inappropriate placement of focal stress, preponderance of falling tones.

Description:

Hesitation and pausing combine to result in short tone units which do not reflect syntactic boundaries and do not always receive natural placement of focal stress. For example, the speakers says:

I think \the | first ↗one | is | near the ↗urban \air

A more natural utterance would certainly have fewer tone units and therefore fewer focal stresses.

The utterance as stated by a native speaker might look more like the example below:

I think the ↗first one | is near the \urban air

Consistent with many Chinese L1 speakers this speaker applies a tone to many of the syllables which results in some irregular lexical stress such as 'peaceful' having both syllables stressed and a lexical tone assigned to each one:

\peace↗ful

Over use of focal stress combined with this speaker's preponderance for falling tones results in speech which sounds staccato:

a lot of \people | \play | to\gether | they \may | they



Transcription:

Let's start the first photograph I think the first one is near the urban air they erm they maybe quite peaceful and very quiet there is suitable for the people they are they like some quiet environment and they are also very green and I think the second photograph is in the city and maybe very crowd and very noisy and a lot of people play together they may they are not a family and they didn't know each other so they make them feel very not comfortable and I think I think the first one is suitable for a group of people maybe some day everybody know each other and the second is just for the just enjoy in their free time

let's start the first photo\graph | I think \the | first \one | is | near the \urban \air | they | erm | they | maybe quite \peaceful | and very \quiet | there is \suitable for the | for the \people they are | they like some quiet en\vironment | and they are also very \green | and I think the second photograph is in the \city | and maybe very \crowd | and very \noisy | and |a lot of \people | \play | to\gether | they \may | they | they are not | they are not a \family | and they didn't know each \other | so they make them feel \very | not com\fortable | and | I \think | I think | that the first one is \suitable for the | for group of \people | er maybe some \day | everybody know each \other | and the \second is | just for the | just enjoy in their free \time

Speaker 02

Spanish | PET

Summary: Extensive use of rising pitch

Description:

The speaker repeats the same pattern of rising pitch throughout the performance. Different intonation patterns are used but the performance is dominated by tone units that end with a steep rising tone followed by a pause as can be seen in the below f0 chart:



The effect of this is to make the speaker sound unsure and tentative.

Transcription:

This is a pharmacy there are three persons two womans and a man er the shop keeper is putting on the hand of the woman some perfume er the man is waiting er for the for the woman er there are a lot of er boxes of perfume and and some stuff and the woman er have a red scarf er a watch on his hand er maybe they are married because the man have a ring on her hand erm the shopkeeper erm have a short hair both of the woman are blonde

Speaker 03

French | FCE

Summary: Preponderance of exaggerated rising tones.

Description:

The speaker uses exaggerated rising tones extensively. This may be interference from the speaker's L1 or a social accent of sorts. Alternatively, it may be the result of the context: rising pitch has been considered to be associated with uncertainty and tentativeness as well as politeness in the face of authority. All these things are consistent with the testing context. The result is a lack of nuance and naturalness to the speaker's tone choice.

Example:

∨OK | so the first ↗picture | I can see a ↗family | probably in the ↗kitchen | so it's umm I think it's
a ↗mother | with some ↗daughters | and they are cooking a ↗pie



Transcription:

OK so the first picture I can see a family probably in the kitchen so it's umm I think it's a mother with some daughters and they are cooking a pie Whereas in the second picture it's umm I think it's in a restaurant a kitchen in a restaurant with some professional so there's a lot of people and adults so the for the first one the think can be difficult it's err, they are small children so maybe they are this is the first time for them to learn how to cook so first of all they need to learn the [indecipherable] to wash their hands before and how to cut or maybe it's the mother can cut where on the other hand the southern picture I think it's some professional or some student about cooking

| \OK | so the first /picture | I can see a /family | probably in the /kitchen | so it's umm I think it's a /mother | with some /daughters | and they are cooking a /pie | Whereas | in the second \picture | it's umm | I think it's in a /restaurant | a kitchen in a /restaurant | with some /professional | so there's a lot of \people | and /adults | so \the | for the first \one | the \think | can be diffi\cult | it's err, they are /small childrens | so \maybe they are | this is the first time for /them | to learn how to /cook | so first of all they need to learn the a/gem | so to wash their hands /before | and how to /cut | or maybe it's the mother can /cut | where on the other /hand | the southern /picture | I think it's some pro\ffessional | or some \student about cooking |

Speaker 04

Spanish | CAE

Summary: Ineffective use of emphatic stress.

Description:

The speaker consistently applies emphatic stress to the same word throughout the performance. She introduces 'flowers' and 'ice' in the first few utterances and when she revisits these words later she stresses them emphatically each time.

The speaker does not take advantage of focal stress to demonstrate later in the performance that this is a given. The result is a sense of repetition.

It may be the case that the speaker is applying emphatic stress to 'ice'. This is perhaps reasonable given the context and how she describes herself as being 'shocked' at the use of ice. However, in this case we would perhaps expect 'ice' to receive a falling pitch.

Transcription:

Erm I think the most difficult materials to use in for artistic for artistic issues are er flowers and ice specially, er the third picture seems to be like er an ice sculpture and I'm really shocked about the result of this sculpture because I think it may be really difficult to try to make er a sculpture with ice and maybe it it seems like even more difficult that trying to to make a figure with wood or stone because I mean ice, can be it can be broken more easily than with [indecipherable] and the hand flowers even the people may think that er doing artistic things with flowers may be easy I don't think it is.

Speaker 05

Vietnamese | FCE

Summary

Inappropriate focal stress, limited rhythm.

Description:

In longer runs she tends to apply a tone to many of the syllables which is consistent with speakers whose L1 is a tone language. When combined with pausing and a lack of connected speech the results is choppy speech lacks a consistent sense of rhythm.



theyjust need to go the supermarket and they can buy

The speaker applies some unusual pitch changes, for example in the following utterance she stresses 'thing' with falling pitch but not with any change in loudness or length.

in the first \picture | is the family go to the \supermarket to | buy the \thing

Transcription:

Er in the first picture is the family go to the supermarket to buy the thing erm and I think in the supermarket is of er because people today is quite busy so they just need to go the supermarket and they can find all of the things they need like from the foods or the souvenirs or some things ah in the second picture is er about the market and erm is very crowded and erm people can buy or sell the food erm erm I think is you if you buy look a free because you can choose anything you want but the quality maybe is not good more than in the supermarket so I prefer to buy the food or something and eat in the supermarket because it's more convenient.

Speaker 06

Chinese | PET

Summary: Short tone units, over use of pitch accents, limited rhythm.

Description:

The speaker applies a tone to a large number of syllables which results in stress appearing to fall very regularly. The result is a sense of choppiness. This is compounded by hesitation and inappropriate pausing which results in there being few long runs and a lack of rhythm.

a family | there in the | maybe in the hotels | they | they put their clothes | put their clothes | on the shelf

Transcription:

From the photograph I can see is there is a family there in the maybe in the hotels they put their clothes put their clothes on the shelf and then the father and his sons and they go sight seeing outside and they maybe they talk about the beautiful places and and the daughter had her to put her clothes on the shelf it is maybe it is maybe in the afternoon they are very friendly and enjoy their holiday.

Speaker 08

Russian | PET

Summary: Inappropriate lexical stress

Description:

The speaker mispronounces 'encyclopaedia' as /,ensək'lopedɪ/, rather than /en,sɑɪklə'pi:diə/ which would be a more accurate pronunciation. The number of syllables is correct but the shift in lexical stress in combination with the different quality of the vowels is likely to influence how easy it is to recognise the words.

Transcription:

Well er this photographs erm I think grandfather and er grand and er child son of I know read some books I think it's er like encyclopaedia this is a big building because is the is big area is big place er they sitting on the red chair arm not its armchair I think erm I think this a new building because a wall don't have a some pictures painting state only a little tv set and er lamp and er I don't know I feel this it's a new building

Speaker 09

Vietnamese | PET

Summary:

Irregular focal stress, limited speech rhythm.

Description

Focal stress is irregularly placed. The interference caused by hesitation and inappropriate pausing results in limited long runs and short tone units. The effect is that there is limited opportunity for there to be a consistent 'beat' to the presentation of focal stress.

Transcription

In this two photographer and erm the first I think that erm this river is in a city and is not a natural and in the second photograph is a natural river and is quite peaceful is not like the first para not like the first photograph because there's a lot of people and erm I think that's erm this the place by people and erm so they can objective a lot of people but erm in the second photograph er I think just the local people visit this erm this place.

Speaker 11

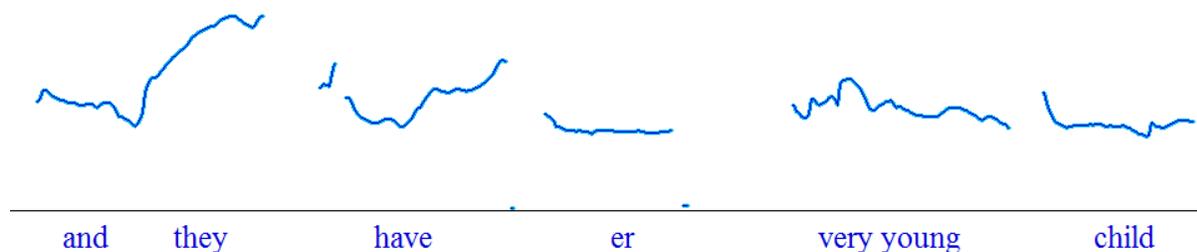
Italian | PET

Summary:

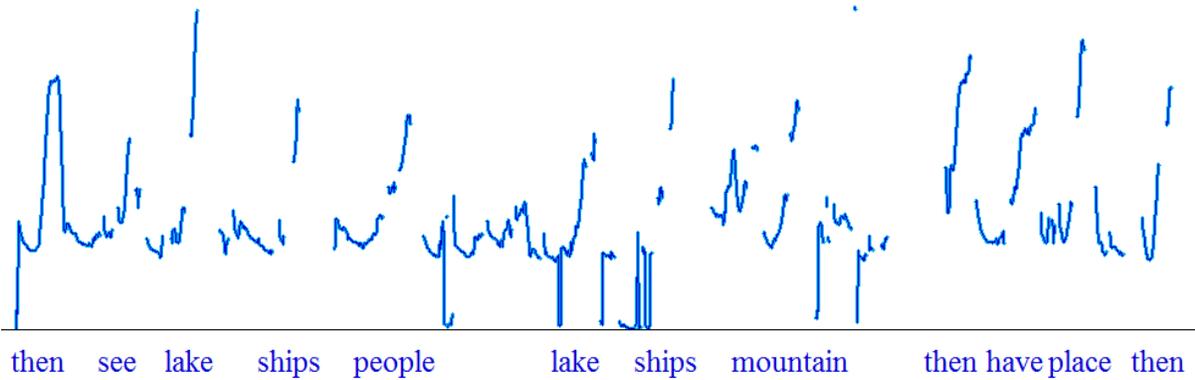
Inappropriate placement of focal stress, overuse and exaggerated rising tone.

Description:

The speaker applies pitch accents to syllables that would not normally be stressed. For example in the following utterance rising tones appears on 'they' and 'have' making them appear to be emphatically stressed.



Steep exaggerated rising tones dominate the performance with the effect that they lose any prosodic meaning.



Transcription

There is a young family erm er he er they are married because I see that he has a wedding ring and they have erm young erm very young child he is one one year old no more and then we can see a lake with several ships and there are people that are looking this lake and the ships and there is also a mountain with several houses then we have a place with flowers and behind this family there is a bar

Speaker 12

Korean | PET

Summary

Narrow pitch range, limited long runs.

Description

The speaker applied pitch movement extensively

For example each word in the utterance “he is smiling to her” has falling tone

Transcription

Is there a couple a man or woman there looks like old couples and he is smiling to her very kindly and friendly and give to her some present like foods with flowers and she was cooking when he give to present to her and it so here is kitchen and some looks like microwave and some bowls and there a sink some spoons cup of tea cup of coffee and also he is wearing jean shirt and

Speaker 14

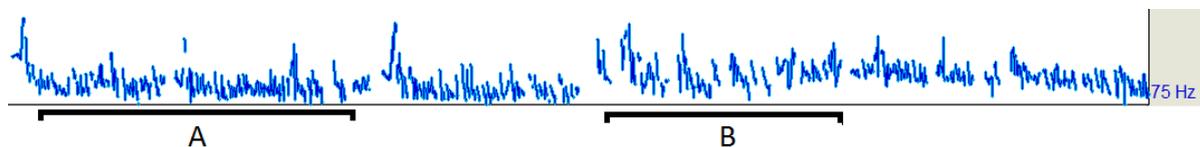
German | CAE

Summary

Narrow pitch range, relatively non salient focal stress.

Description

The speech is fast with a narrow prominence gradient. This has the effect of making the speech sound monotonous. This is common L1 interference of German speakers. This can be seen in the f0 chart below where in section A there are few gaps and a narrow pitch range but at section B the speaker starts to speak more slowly and pause more normally, also the pitch range broadens resulting in much more natural sounding speech.



Transcription

Yes er I would like to start with the second picture when the children I see the children playing football and I think it's really good opportunity in life to erm learn so many things er I'd like to talk about social competence probably somebody some days they gona win somedays they gona lose so there's a lot of things to learn already when children or when we are young but the most important thing is just to have a good time together feel as a children so children shouldn't go all the day in school and learn the so many things so that's really important thing let the children do what they like to do and sport is very important in the same age compared probably to the second one er to the first picture they are playing together in the family so probably for me it looks like

their parents playing monopoly together and just enjoy the evening and to to to get the feeling how life can be together in the family to really have this relationship from the children to the parents erm and really to learn more about life on the one hand but also to learn or how in the evening or how we could spend the time together.

Appendix F: Final Audio Rationales

Speaker	Lexical Stress	Rhythm	Intonation
1			Initially monotonic; preponderance of steep falling tones; short intonational phrases
2	Lexical stress errors, e.g. "perfume"		Preponderance of steep rising tones; short intonational phrases
3			Preponderance of steep rising tones
4		Inappropriate rhythm (caused by hesitation and inappropriate vowel duration)	Monotonic; rising intonation at the end of many phrases
5	Lexical stress errors, e.g. "CONvenient"	Limited rhythm (lack of connected speech and lack of reduction)	Monotonic
6		Limited rhythm (caused by inappropriate pausing and hesitation);	Short tone units but appropriate intonation within tone units
8	Lexical stress errors: "encyclopedia", "grandfather", "TV"	Excessive hesitation	
9	Lexical stress errors on "photographer", "photograph"	Limited speech rhythm (caused by hesitation, and lack of relative salience)	Monotonic; predominance of rising tones
10		Periods of limited rhythm (caused by lack of linking and lack of reduction);	Lack of continuation intonation; excessive rising intonation at the end of short intonational units; monotone
11		Periods of limited speech rhythm (caused by hesitation, inappropriate pausing, limited linking and reduction, addition of extra vowels at the end of words)	Preponderance of rising tones (some steeply rising)
12		Limited rhythm (caused by hesitation, pausing)	Preponderance of level tones; too many rises at the end of phrases
14			Narrow pitch range and limited pitch excursions

Appendix G: Standardised Scores

Non-Experienced Raters															
Score	NR_01	NR_02	NR_03	NR_04	NR_05	NR_06	NR_07	NR_08	NR_09	NR_10	NR_11	NR_12	NR_13	NR_14	NR_15
1	-0.18	-0.18	-0.18	-0.18	5.29	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18
2	-0.70	-0.70	2.43	0.35	0.35	0.35	-0.70	0.35	0.35	0.35	2.43	-0.70	0.35	-0.70	0.35
3	-1.25	0.31	1.10	0.31	0.31	0.31	-0.47	0.31	1.10	0.31	1.88	-0.47	1.88	-0.47	0.31
4	1.33	0.62	-0.09	-0.81	-0.09	0.62	1.33	1.33	0.62	1.33	-1.52	-0.09	-0.81	1.33	-0.81
5	1.79	0.57	-0.65	-1.87	0.57	-1.87	-0.65	0.57	0.57	-1.87	-0.65	0.57	-0.65	0.57	0.57
6	-0.55	-0.55	-1.17	-0.55	-0.55	-0.55	0.68	0.06	-1.17	0.68	-1.17	0.68	-0.55	0.06	-0.55
7	-0.57	-0.57	0.21	1.76	-1.35	0.21	-0.57	-0.57	0.21	-0.57	-1.35	1.76	0.21	-1.35	0.21
8	-0.35	-0.35	-1.24	-0.35	1.41	-0.35	-0.35	-1.24	-0.35	-0.35	1.41	-1.24	-0.35	1.41	-1.24
9	1.15	1.15	0.00	1.15	-1.15	1.15	0.00	-1.15	-1.15	-1.15	0.00	-1.15	0.00	-1.15	2.30
Experienced Raters															
Score	ER_01	ER_02	ER_03	ER_04	ER_05	ER_06	ER_07	ER_08	ER_09	ER_10	ER_11	ER_12	ER_13	ER_14	ER_15
1	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18
2	-0.70	-0.70	-0.70	2.43	-0.70	-0.70	-0.70	-0.70	1.39	-0.70	-0.70	-0.70	0.35	-0.70	-0.70
3	-1.25	-1.25	-0.47	0.31	-0.47	-1.25	-0.47	1.88	-0.47	-0.47	-1.25	-1.25	1.10	1.10	-1.25
4	-0.81	-1.52	1.33	-0.09	-0.09	-0.81	-0.09	-0.81	-0.09	1.33	-1.52	-1.52	-0.09	-0.81	1.33
5	0.57	-0.65	0.57	0.57	-0.65	1.79	0.57	0.57	0.57	0.57	0.57	0.57	-1.87	-0.65	-0.65
6	-0.55	1.29	-1.17	-1.17	1.91	0.68	0.68	-0.55	0.06	0.06	2.52	0.06	-1.17	1.91	0.68
7	0.99	1.76	0.21	-0.57	-0.57	-0.57	0.21	-1.35	0.21	-1.35	0.21	1.76	1.76	-0.57	0.21
8	2.30	0.53	0.53	0.53	-0.35	1.41	0.53	1.41	-0.35	0.53	-1.24	0.53	-1.24	-0.35	-1.24
9	0.00	0.00	0.00	-1.15	0.00	0.00	-1.15	0.00	-1.15	0.00	1.15	1.15	1.15	-1.15	1.15

Appendix H: Facets Output

All Raters Examinee Score Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea PtExp		Nu Examinees
253	30	8.43	8.47	3.77	.28	.94	-.1	.86	-.4	1.15	.50	.36	9 Speaker_10
240	30	8.00	8.03	2.95	.23	1.06	.3	1.08	.3	.92	.46	.43	12 Speaker_14
219	30	7.30	7.33	2.00	.20	1.14	.5	1.19	.7	.77	.08	.49	3 Speaker_03
188	30	6.27	6.35	1.00	.16	.81	-.6	.80	-.7	1.13	.49	.56	4 Speaker_04
158	30	5.27	5.32	.27	.15	.80	-.8	.82	-.6	1.11	.68	.60	10 Speaker_11
154	30	5.13	5.17	.18	.15	1.01	.1	1.02	.1	.91	.42	.60	5 Speaker_05
146	30	4.87	4.87	.00	.15	1.44	1.6	1.36	1.4	.51	.55	.61	2 Speaker_02
144	30	4.80	4.79	-.05	.15	1.16	.7	1.12	.5	1.01	.63	.61	7 Speaker_08
127	30	4.23	4.17	-.44	.16	.83	-.6	.81	-.7	1.32	.77	.60	11 Speaker_12
122	30	4.07	3.99	-.57	.16	.58	-1.9	.60	-1.7	1.41	.65	.60	6 Speaker_06
118	30	3.93	3.85	-.67	.16	.91	-.2	.90	-.3	1.16	.65	.59	8 Speaker_09
114	30	3.80	3.71	-.77	.16	.89	-.3	.97	.0	1.00	.65	.58	1 Speaker_01
165.3	30.0	5.51	5.50	.64	.18	.96	-.1	.96	-.1		.54		Mean (Count: 12)
46.4	.0	1.55	1.59	1.43	.04	.21	.9	.20	.8		.17		S.D. (Population)
48.5	.0	1.62	1.66	1.50	.04	.22	.9	.21	.8		.18		S.D. (Sample)

Model, Populn: RMSE .18 Adj (True) S.D. 1.42 Separation 7.89 Strata 10.85 Reliability .98
 Model, Sample: RMSE .18 Adj (True) S.D. 1.49 Separation 8.25 Strata 11.33 Reliability .99
 Model, Fixed (all same) chi-square: 521.3 d.f.: 11 significance (probability): .00
 Model, Random (normal) chi-square: 10.8 d.f.: 10 significance (probability): .38

Non-Experienced Rater Examinee Measurement Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu Examinees
124	15	8.27	8.30	2.87	.32	1.03	.1	1.01	.1	1.11	.37	.35	9 Speaker_10
117	15	7.80	7.82	2.25	.27	.84	-.3	.86	-.2	1.23	.62	.40	12 Speaker_14
113	15	7.53	7.56	1.97	.26	.89	-.1	.98	.0	.92	.23	.42	3 Speaker_03
88	15	5.87	5.91	.72	.20	.84	-.3	.83	-.3	1.11	.35	.52	4 Speaker_04
73	15	4.87	4.83	.13	.20	.88	-.2	.87	-.3	1.06	.50	.52	10 Speaker_11
72	15	4.80	4.76	-.09	.20	.85	-.3	.85	-.3	1.08	.47	.51	5 Speaker_05
69	15	4.60	4.55	-.03	.20	1.20	.6	1.17	.5	.89	.43	.50	7 Speaker_08
60	15	4.00	3.94	-.42	.22	.80	-.4	.75	-.6	1.12	.39	.46	6 Speaker_06
60	15	4.00	3.94	-.42	.22	.96	.0	.89	-.1	1.23	.59	.46	11 Speaker_12
57	15	3.80	3.75	-.57	.23	1.56	1.4	1.34	.9	.68	.18	.45	2 Speaker_02
50	15	3.33	3.30	-.96	.25	1.05	.2	1.04	.2	1.03	.47	.41	8 Speaker_09
48	15	3.20	3.17	-1.09	.26	.96	.0	.96	.0	1.06	.71	.40	1 Speaker_01
77.6	15.0	5.17	5.15	.38	.24	.99	.1	.96	.0		.44		Mean (Count: 12)
25.6	.0	1.71	1.74	1.25	.04	.20	.5	.16	.4		.15		S.D. (Population)
26.8	.0	1.78	1.82	1.31	.04	.21	.5	.16	.4		.15		S.D. (Sample)

Model, Populn: RMSE .24 Adj (True) S.D. 1.23 Separation 5.16 Strata 7.21 Reliability .96
 Model, Sample: RMSE .24 Adj (True) S.D. 1.29 Separation 5.40 Strata 7.53 Reliability .97
 Model, Fixed (all same) chi-square: 255.0 d.f.: 11 significance (probability): .00
 Model, Random (normal) chi-square: 10.5 d.f.: 10 significance (probability): .39

Experienced Rater Examinee Measurement Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu Examinees
129	15	8.60	8.63	4.99	.51	.59	-1.2	.55	-1.2	1.50	.65	.39	9 Speaker_10
123	15	8.20	8.22	3.71	.42	1.45	1.1	1.48	1.2	.43	.24	.45	12 Speaker_14
106	15	7.07	7.12	1.43	.32	.99	.0	.98	.0	1.01	.26	.56	3 Speaker_03
100	15	6.67	6.73	.88	.29	.80	-.4	.84	-.3	1.14	.46	.59	4 Speaker_04
89	15	5.93	6.03	-.05	.26	1.22	.6	1.21	.6	.69	.54	.64	2 Speaker_02
85	15	5.67	5.77	-.21	.25	.81	-.4	.93	.0	1.08	.78	.65	10 Speaker_11
82	15	5.47	5.56	-.39	.25	1.40	1.1	1.41	1.1	.53	.26	.66	5 Speaker_05
75	15	5.00	5.05	-.80	.24	1.14	.5	1.09	.3	.98	.80	.67	7 Speaker_08
68	15	4.53	4.52	-1.21	.24	.96	.0	.96	.0	1.06	.62	.68	8 Speaker_09
67	15	4.47	4.44	-1.26	.24	.73	-.7	.77	-.5	1.43	.93	.68	11 Speaker_12
66	15	4.40	4.36	-1.32	.24	1.13	.4	1.18	.5	.73	.45	.68	1 Speaker_01
62	15	4.13	4.06	-1.56	.25	.19	-3.4	.21	-3.3	2.00	.95	.67	6 Speaker_06
87.7	15.0	5.84	5.87	.36	.29	.95	-.2	.97	-.1		.58		Mean (Count: 12)
21.6	.0	1.44	1.46	2.00	.08	.34	1.2	.34	1.2		.24		S.D. (Population)
22.5	.0	1.50	1.53	2.09	.09	.36	1.3	.35	1.2		.25		S.D. (Sample)

Model, Populn: RMSE .30 Adj (True) S.D. 1.98 Separation 6.49 Strata 8.98 Reliability .98
 Model, Sample: RMSE .30 Adj (True) S.D. 2.07 Separation 6.78 Strata 9.38 Reliability .98
 Model, Fixed (all same) chi-square: 322.0 d.f.: 11 significance (probability): .00
 Model, Random (normal) chi-square: 10.6 d.f.: 10 significance (probability): .39

All Raters Measurement Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Obs %	Agree. Exp %	Nu Raters	
53	12	4.42	4.20	1.06	.28	.47	-1.4	.50	-1.3	1.52	.95	.88	16.7	17.3	19	ER_04
51	12	4.25	4.39	.94	.29	.58	-1.0	.65	-.8	1.42	.94	.88	26.2	22.1	3	NR_03
53	12	4.42	4.65	.78	.28	.53	-1.2	.63	-.8	1.21	.92	.88	23.8	23.2	8	NR_08
55	12	4.58	4.90	.62	.27	.25	-2.5	.25	-2.5	1.69	.95	.87	22.6	24.1	9	NR_09
59	12	4.92	4.91	.62	.27	1.22	.6	1.17	.5	.72	.77	.86	22.0	21.3	24	ER_09
56	12	4.67	5.02	.55	.27	1.43	1.0	1.26	.7	.70	.89	.87	23.8	24.5	11	NR_11
57	12	4.75	5.14	.48	.27	1.38	.9	1.39	.9	.41	.84	.87	23.8	24.8	5	NR_05
57	12	4.75	5.14	.48	.27	.89	-.1	1.06	.2	1.02	.80	.87	22.6	24.8	10	NR_10
63	12	5.25	5.36	.34	.26	1.03	.1	1.07	.3	.96	.76	.84	20.8	23.3	29	ER_14
60	12	5.00	5.48	.26	.26	1.31	.8	1.37	.9	.67	.80	.86	27.4	25.3	13	NR_13
65	12	5.42	5.57	.21	.26	1.15	.4	1.11	.4	.82	.82	.83	17.9	24.1	23	ER_08
65	12	5.42	5.57	.21	.26	1.30	.8	1.19	.5	1.01	.85	.83	23.8	24.1	28	ER_13
66	12	5.50	5.67	.14	.26	.37	-2.0	.36	-2.1	1.73	.93	.83	26.8	24.4	25	ER_10
68	12	5.67	5.86	.01	.26	.51	-1.3	.46	-1.6	1.64	.90	.82	25.0	24.8	18	ER_03
69	12	5.75	5.95	-.05	.26	.61	-1.0	.67	-.8	1.30	.80	.82	25.0	25.0	22	ER_07
65	12	5.42	5.98	-.07	.26	.43	-1.7	.40	-1.8	1.59	.89	.83	24.4	25.1	14	NR_14
70	12	5.83	6.04	-.12	.26	.75	-.5	.87	-.2	1.19	.79	.81	23.8	25.2	20	ER_05
66	12	5.50	6.07	-.14	.26	1.57	1.3	1.44	1.1	.72	.82	.83	26.2	24.8	6	NR_06
66	12	5.50	6.07	-.14	.26	.80	-.4	.80	-.4	1.31	.84	.83	24.4	24.8	7	NR_07
71	12	5.92	6.13	-.18	.26	.77	-.5	.74	-.5	1.41	.85	.81	27.4	25.2	30	ER_15
67	12	5.58	6.16	-.20	.26	1.79	1.7	1.78	1.7	.02	.71	.82	18.5	24.6	2	NR_02
67	12	5.58	6.16	-.20	.26	1.11	.3	1.45	1.1	.53	.64	.82	22.0	24.6	12	NR_12
69	12	5.75	6.32	-.34	.26	2.16	2.3	2.03	2.1	-.15	.74	.82	22.0	23.9	15	NR_15
70	12	5.83	6.40	-.40	.26	.45	-1.6	.47	-1.5	1.62	.93	.81	25.6	23.5	1	NR_01
72	12	6.00	6.55	-.53	.26	1.93	1.9	1.66	1.5	.33	.72	.80	16.7	22.6	4	NR_04
77	12	6.42	6.61	-.59	.27	1.27	.7	1.44	1.0	.39	.64	.78	17.3	24.8	21	ER_06
78	12	6.50	6.69	-.66	.27	.54	-1.2	.62	-.9	1.39	.87	.77	26.8	24.6	26	ER_11
82	12	6.83	6.99	-.97	.29	.99	.1	.92	.0	.96	.70	.75	23.8	23.3	16	ER_01
82	12	6.83	6.99	-.97	.29	.39	-1.6	.50	-1.2	1.45	.83	.75	25.6	23.3	17	ER_02
84	12	7.00	7.14	-1.13	.30	.51	-1.1	.59	-.8	1.35	.81	.74	25.0	22.3	27	ER_12
66.1	12.0	5.51	5.80	.00	.27	.95	-.2	.96	-.1		.82					Mean (Count: 30)
8.7	.0	.72	.76	.55	.01	.50	1.3	.46	1.2		.09					S.D. (Population)
8.8	.0	.74	.77	.56	.01	.50	1.3	.47	1.2		.09					S.D. (Sample)

Model, Populn: RMSE .27 Adj (True) S.D. .48 Separation 1.81 Strata 2.75 Reliability (not inter-rater) .77
 Model, Sample: RMSE .27 Adj (True) S.D. .49 Separation 1.85 Strata 2.80 Reliability (not inter-rater) .77
 Model, Fixed (all same) chi-square: 117.7 d.f.: 29 significance (probability): .00
 Model, Random (normal) chi-square: 23.5 d.f.: 28 significance (probability): .71
 Inter-Rater agreement opportunities: 2520 Exact agreements: 586 = 23.3% Expected: 601.2 = 23.9%

Non-Experienced Raters Measurement Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Obs %	Agree. Exp %	Nu Raters
51	12	4.25	4.01	.75	.28	.50	-1.2	.53	-1.1	1.55	.95	.87	26.2	21.8	3 NR_03
53	12	4.42	4.24	.60	.27	.54	-1.1	.61	-.9	1.21	.92	.87	23.8	22.9	8 NR_08
55	12	4.58	4.47	.46	.27	.25	-2.4	.26	-2.4	1.67	.96	.87	22.6	23.8	9 NR_09
56	12	4.67	4.59	.38	.27	1.38	.9	1.15	.4	.78	.90	.87	23.8	24.1	11 NR_11
57	12	4.75	4.71	.32	.26	.98	.0	1.03	.2	.76	.90	.87	23.8	24.4	5 NR_05
57	12	4.75	4.71	.32	.26	.70	-.6	.75	-.5	1.26	.86	.87	22.6	24.4	10 NR_10
60	12	5.00	5.07	.12	.26	1.15	.4	1.15	.4	.92	.83	.86	27.4	24.9	13 NR_13
65	12	5.42	5.66	-.20	.25	.35	-2.0	.32	-2.2	1.51	.91	.84	24.4	24.6	14 NR_14
66	12	5.50	5.77	-.26	.25	1.22	.6	1.13	.4	1.12	.85	.84	26.2	24.4	6 NR_06
66	12	5.50	5.77	-.26	.25	.81	-.3	.96	.0	1.14	.82	.84	24.4	24.4	7 NR_07
67	12	5.58	5.87	-.32	.25	1.86	1.8	1.76	1.6	.24	.70	.84	18.5	24.1	2 NR_02
67	12	5.58	5.87	-.32	.25	1.04	.2	1.36	.9	.57	.67	.84	22.0	24.1	12 NR_12
69	12	5.75	6.08	-.44	.25	1.95	2.0	1.80	1.7	.21	.75	.83	22.0	23.5	15 NR_15
70	12	5.83	6.18	-.50	.25	.54	-1.2	.50	-1.4	1.55	.91	.82	25.6	23.2	1 NR_01
72	12	6.00	6.37	-.62	.25	1.25	.7	1.14	.4	.97	.79	.81	16.7	22.3	4 NR_04
62.1	12.0	5.17	5.29	.00	.26	.97	-.1	.96	-.1		.85				Mean (Count: 15)
6.6	.0	.55	.75	.42	.01	.50	1.3	.46	1.2		.09				S.D. (Population)
6.8	.0	.57	.78	.44	.01	.51	1.3	.47	1.3		.09				S.D. (Sample)

Model, Populn: RMSE .26 Adj (True) S.D. .34 Separation 1.32 Strata 2.10 Reliability (not inter-rater) .64
 Model, Sample: RMSE .26 Adj (True) S.D. .36 Separation 1.40 Strata 2.19 Reliability (not inter-rater) .66
 Model, Fixed (all same) chi-square: 40.1 d.f.: 14 significance (probability): .00
 Model, Random (normal) chi-square: 10.5 d.f.: 13 significance (probability): .65
 Inter-Rater agreement opportunities: 1260 Exact agreements: 294 = 23.3% Expected: 299.7 = 23.8%

Experienced Raters Measurement Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Obs %	Agree. Exp %	Nu Raters
53	12	4.42	4.56	1.53	.31	.45	-1.5	.47	-1.4	1.65	.94	.90	16.7	18.1	19 ER_04
59	12	4.92	5.29	.97	.30	1.21	.6	1.18	.5	.73	.80	.88	22.0	23.0	24 ER_09
63	12	5.25	5.70	.63	.29	.99	.0	1.03	.2	1.06	.80	.86	20.8	25.5	29 ER_14
65	12	5.42	5.88	.46	.29	1.72	1.5	1.73	1.6	.17	.78	.85	17.9	26.4	23 ER_08
65	12	5.42	5.88	.46	.29	1.32	.8	1.26	.7	.77	.86	.85	23.8	26.4	28 ER_13
66	12	5.50	5.96	.38	.29	.60	-1.0	.59	-1.1	1.46	.90	.85	26.8	26.8	25 ER_10
68	12	5.67	6.13	.21	.29	.44	-1.6	.47	-1.5	1.60	.90	.84	25.0	27.4	18 ER_03
69	12	5.75	6.21	.12	.29	.70	-.6	.69	-.7	1.29	.81	.83	25.0	27.6	22 ER_07
70	12	5.83	6.28	.04	.29	.95	.0	1.05	.2	.99	.80	.83	23.8	27.8	20 ER_05
71	12	5.92	6.36	-.05	.29	.88	-.1	.84	-.3	1.19	.87	.82	27.4	27.9	30 ER_15
77	12	6.42	6.78	-.58	.31	1.45	1.0	1.79	1.7	.15	.66	.80	17.3	27.2	21 ER_06
78	12	6.50	6.85	-.68	.31	.76	-.4	.81	-.3	1.19	.86	.80	26.8	26.8	26 ER_11
82	12	6.83	7.13	-1.09	.33	1.30	.7	1.20	.5	.76	.70	.78	23.8	24.9	16 ER_01
82	12	6.83	7.13	-1.09	.33	.62	-.8	.66	-.7	1.30	.81	.78	25.6	24.9	17 ER_02
84	12	7.00	7.28	-1.32	.35	.76	-.4	.72	-.5	1.18	.81	.77	25.0	23.5	27 ER_12
70.1	12.0	5.84	6.23	.00	.30	.94	-.1	.97	-.1		.82				Mean (Count: 15)
8.7	.0	.72	.72	.78	.02	.37	.9	.40	1.0		.07				S.D. (Population)
9.0	.0	.75	.74	.81	.02	.38	1.0	.41	1.0		.07				S.D. (Sample)
Model, Populn: RMSE .31 Adj (True) S.D. .72 Separation 2.36 Strata 3.48 Reliability (not inter-rater) .85 Model, Sample: RMSE .31 Adj (True) S.D. .75 Separation 2.46 Strata 3.61 Reliability (not inter-rater) .86 Model, Fixed (all same) chi-square: 90.0 d.f.: 14 significance (probability): .00 Model, Random (normal) chi-square: 12.2 d.f.: 13 significance (probability): .51 Inter-Rater agreement opportunities: 1260 Exact agreements: 292 = 23.2% Expected: 322.8 = 25.6%															

Appendix I: Raw Scores

		Speaker											
		01	02	03	04	05	06	08	09	10	11	12	14
Non-Experienced Raters	NR_01	4	6	8	7	4	5	4	4	9	5	5	9
	NR_02	3	8	9	4	3	5	6	4	9	5	4	7
	NR_03	2	3	7	5	2	4	3	2	9	4	3	7
	NR_04	3	2	7	7	7	6	8	3	9	7	4	9
	NR_05	1	3	8	4	5	5	6	2	8	4	3	8
	NR_06	4	3	8	7	6	4	4	2	9	3	7	9
	NR_07	4	3	6	7	6	4	4	6	9	5	4	8
	NR_08	2	4	6	5	4	3	4	4	7	5	3	6
	NR_09	2	4	7	5	4	3	3	3	8	5	4	7
	NR_10	2	4	8	6	4	4	6	3	7	4	3	6
	NR_11	3	3	8	8	5	3	3	2	9	2	2	8
	NR_12	4	4	6	7	5	6	6	5	7	7	3	7
	NR_13	4	3	8	5	7	3	3	2	7	6	3	9
	NR_14	4	4	8	6	5	3	6	4	8	4	5	8
	NR_15	6	3	9	5	5	2	3	4	9	7	7	9
Experienced Raters	ER_01	6	8	8	7	5	5	7	4	9	8	7	8
	ER_02	5	6	7	8	7	6	7	6	9	6	7	8
	ER_03	4	7	8	7	5	4	4	4	9	5	3	8
	ER_04	3	5	7	5	4	2	3	2	8	4	2	8
	ER_05	3	5	6	6	7	4	6	6	8	6	4	9
	ER_06	6	8	8	6	4	5	5	5	9	8	6	7
	ER_07	3	6	6	7	5	4	7	5	8	6	4	8
	ER_08	5	3	8	8	3	3	4	3	8	6	5	9
	ER_09	5	5	6	7	6	3	2	4	8	4	2	7
	ER_10	4	5	8	6	5	4	6	4	9	4	3	8
	ER_11	5	6	7	6	6	6	7	6	9	5	6	9
	ER_12	5	7	7	8	6	6	7	5	9	8	7	9
	ER_13	2	7	7	7	7	3	3	4	9	4	3	9
	ER_14	6	6	6	6	6	3	3	3	8	5	4	7
	ER_15	4	5	7	6	6	4	4	7	9	6	4	9

Appendix J: Categories and Codes

Group 1: Suprasegmental Categories

Intonation		Experienced Raters Count	Non-Experienced Raters Count
Broad	Intonation	120	42
	Inflection	6	2
	Cadence	5	7
	Tone of voice	0	1
Descriptive	Bored	0	1
	Boring	0	2
	Bouncing	0	1
	Lilt	0	3
Technical	Focal Stress	56	9
	Monotone	15	13
	Pitch	4	11
	Tone	4	17
	Emphasis	2	10

Rhythm		Experienced Raters Count	Non-Experienced Raters Count
Broad	Rhythm	53	18
	Musical	9	1
	Sing song	0	2
Fragmentation	Hesitation	41	24
	Pausing	24	46
	Staccato	16	1
	Halting	5	0
	Stilted	6	10
	Stopping starting	5	18
	Choppy	4	5
	Disjointed	4	2
	Fragmented	2	2
	Machine-gun delivery	1	0
	Jerky	1	0
	Segmented	1	0
	Stalling	1	0
	Broken(-up)	0	12
	Jittery	0	2
	Joined-up	0	3

	Stutter	0	5
Technical	Linking	17	6
	Weak forms	13	0
	Connected speech	7	0

Lexical Stress		Experienced Raters Count	Non-Experienced Raters Count
Lexical Stress	Lexical Stress	30	12

Group 2: Other Categories

Other Categories		Experienced Raters Count	Non-Experienced Raters Count
Accent	Accent	18	69
	Accent Familiarity	34	14
	L1 Interference	23	19
Infer Test Taker Characteristics	Calm	0	1
	Confidence	16	55
	Infer Test Taker Characteristics	15	10
	Nervous	8	18
Strain	Annoying	0	1
	Concentration	6	10
	Effort	1	6
	Frustration	0	5
	Irritation	9	2
	Strain	20	1
Natural	Natural	32	19

Appendix K: Integrated Display

Category	Rhythm	Intonation	Lexical Stress
B1	<p>"very stilted, lots of pauses"</p> <p>"very hesitant, very slow delivery"</p> <p>"he seemed to stagger it with many pauses [...]. several pauses interrupted the natural flow of his speech"</p> <p>"there was a bit of strain on me"</p>	<p>"quite a flat intonation really [...] no rise or fall in intonation"</p> <p>"quite monotonous"</p> <p>"it was that flat intonation that really made it hard"</p> <p>"the emphasis was on the words you wouldn't expect emphasis to be on"</p>	<p>"some words were not clear... the word stress was wrong on perfume"</p> <p>"word stress is not particularly good"</p>
B1+	<p>"it required concentration to listen to... a little bit broken"</p> <p>"there's almost a sort of staccato, little breaks [...] it's sort of quite jittery"</p> <p>"I think the hesitation makes it quite difficult to follow [...] it's choppy"</p>	<p>"intonation... flat... there's no really sense of variation in tone patterns which is again probably first language influence there... it's just boring... monotone"</p> <p>"it's all monotone"</p>	
B2	<p>"fragmented and disrupted by the pauses"</p> <p>"quite a bit of hesitancy and fragmentation"</p>	<p>"unnatural rising intonation [...] rather listy, rather than connected speech"</p> <p>"uncertain about intonation to convey meaning"</p> <p>"rather singsong intonation, very unnatural"</p> <p>"there's no attempt at emphasising really important words so it does make it difficult"</p>	<p>"the internal stress of the words makes some of them more difficult to recognise"</p> <p>"occasional inappropriate stress"</p>
B2+	<p>"a little bit of hesitancy but she always managed to fill it in with an appropriate connective"</p> <p>"quite a bit of hesitation er er... little gaps"</p> <p>"it's a bit choppy but I still think she's obviously used to native English speech patterns because she can follow them"</p>	<p>"there was kind of an attempt to use intonation"</p> <p>"he was using weak forms very naturally "</p> <p>"sentence stress is generally appropriate"</p> <p>"she was emphasising some of the important words"</p>	

C1	<p>"it was a bit choppy, but still it wasn't monotonous so I found listening to her was quite pleasant - it had a nice rhythm going on a nice cadence"</p> <p>"she's breaking it up in quite a natural way"</p>	<p>"very natural intonation"</p> <p>"she had this you know sort of French intonation going up etcetera"</p> <p>"he's using intonation to add more meaning to his utterances [...] so his intonation embellishes the sounds and sentence"</p> <p>"in some places the sentence stress helped us, but he didn't have a very wide range"</p>	<p>"word stresses is generally appropriate"</p> <p>"word stress is accurately placed on the whole"</p>
C1+	<p>"she pauses in just the right places, so she's really got the rhythm of the language down to a tee"</p> <p>"hardly hesitating, maybe there was just a little"</p>	<p>"the inflection is almost like an English person"</p> <p>"very nice stress, very nice intonation "</p> <p>"it's more of a natural cadence"</p> <p>"he was using weak forms very naturally "</p>	<p>"there's very little incorrect word stress"</p> <p>"the word stress was correct"</p>