Isaacs, T., Trofimovich, P., Yu, G., & Munoz Chereau, B. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS Pronunciation scale. British Council.

IELTS

IELTS Research Reports Online Series

ISSN 2201-2982
Reference: 2015/4

Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS Pronunciation scale

Author: *Talia Isaacs, University of Bristol, UK; Pavel Trofimovich, Concordia University, Canada; Guoxing Yu and Bernardita Muñoz Chereau, University of Bristol, UK*

Grant awarded: *Round 17, 2011*

Keywords: *"IELTS Pronunciation scale, Speaking test, comprehensibility, lexicogrammatical measures, examiner ratings, phonological features, mixed methods"*

Abstract

The goal of this study is to identify the linguistic factors that most efficiently distinguish between upper levels of the IELTS Pronunciation scale. Analyses of test-taker speaking performance, coupled with IELTS examiners' ratings of discrete elements and qualitative comments, reveal ways of increasing the transparency of rating scale descriptors for IELTS examiners.

Following the expansion of the IELTS Pronunciation scale from four to nine band levels, the goal of this study is to identify the linguistic factors that most efficiently distinguish between upper levels of the revised IELTS pronunciation scale.

The study additionally aims to identify the trait-relevant variables that inform raters' pronunciation scoring decisions, particularly as they pertain to the 'comprehensible speech' criterion described in the *IELTS Handbook* (IELTS, 2007) and to relate these back to existing rating scale descriptors.

Speech samples of 80 test-takers performing the IELTS long-turn speaking task were rated by eight accredited IELTS examiners for numerous discrete measures shown to relate to the comprehensibility construct, including segmental, prosodic, fluency, and lexicogrammatical measures. These variables, rated on separate semantic-differential scales, were included as predictors in two discriminant analyses, with Cambridge English pre-rated IELTS overall Speaking scores and scores on the Pronunciation subscale used as the grouping variables. Statistical outcomes were then triangulated with the IELTS examiners' focus group data on their use of the IELTS Pronunciation scale levels and the criteria most relevant to their scoring decisions.

Results suggest the need for greater precision in the terminology used in the IELTS Pronunciation subscale to foster more consistent interpretation among raters. In particular, descriptors that were solely distinguished from adjacent bands by stating that the test-taker has achieved all pronunciation features of the lower band but not all those specified in the higher band had poor prediction value and were cumbersome for examiners to use, revealing the need for specific pronunciation features to be delineated at those levels of the scale.

Publishing details

Published by the IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia © 2015.
This online series succeeds *IELTS Research Reports Volumes 1–13*, published 1998–2012 in print and on CD.

AUTHOR BIODATA

Talia Isaacs

Talia Isaacs is a Senior Lecturer in Education at the University of Bristol. She is director of the University of Bristol Second Language Speech Lab, funded through a Marie Curie EU grant (http://www.bris.ac.uk/speech-lab), and co-coordinator of the Centre for Assessment and Evaluation Research (CAER). Her research centres on second language (L2) aural/oral assessment, with a focus on the development and validation of rating scales, the alignment between rater perceptions and L2 speech productions, and oral communication breakdowns and strategies in workplace and academic settings. Talia is an Expert Member of the European Association for Language Testing and Assessment, a founding member of the Canadian Association of Language Assessment, and serves on the Editorial Boards of *Language Assessment Quarterly, Language Testing*, and *The Journal of Second Language Pronunciation*. In addition to her graduate teaching at Bristol, she regularly conducts assessment literacy training for educators within the university and beyond.

Pavel Trofimovich

Pavel Trofimovich is an Associate Professor at the Department of Education's Applied Linguistics Program at Concordia University, Canada. His research focuses on cognitive aspects of L2 processing, L2 phonology, sociolinguistic aspects of L2 acquisition, and teaching L2 pronunciation. Pavel is co-author of two volumes on priming methods in applied linguistics research and is a recipient of the Paul Pimsleur Award for Research in Foreign Language Education along with his Concordia colleagues. He has served as Principal Investigator and Co-Applicant on numerous grants funded by the Social Science and Humanities Research Council of Canada and the Fonds Québécois de la Recherche sur la Société et la Culture on various aspects of L2 pronunciation development and the interaction of classroom input with learner attention. He currently serves as Editor of *Language Learning* and on the Editorial Boards of *Language Learning and Technology* and *The Journal of Second Language Pronunciation*.

Guoxing Yu

Guoxing Yu is a Reader in Language Education and Assessment and Coordinator of Doctor of Education in Applied Linguistics program at the University of Bristol. His main research efforts straddle across: language assessment, the role of language in assessment, assessment of school effectiveness and learning power. He has directed or co-directed several funded research projects and has published in academic journals including *Applied Linguistics, Assessing Writing, Assessment in Education, Educational Research, Language Assessment Quarterly* and *Language Testing*. He was the Guest Editor of the special issue on integrated writing assessment (2013) for *Language Assessment Quarterly*; and the special issue on English Language Assessment in China: Policies, Practices and Impacts (2014) for Assessment in Education (with Prof Jin Yan, Shanghai Jiaotong University),. Dr Yu is an Executive Editor of Assessment in Education, and serves on Editorial Boards of *Language Testing, Language Assessment Quarterly, Assessing Writing* and *Language Testing in Asia*.

Bernardita Muñoz Chereau

Bernardita Muñoz Chereau holds a degree in Psychology from the Catholic University of Chile, a Masters in Education at the University of London, and a PhD in Education at the University of Bristol. Her doctoral work focused on Chilean secondary schools' interpretation of examination results for accountability purposes by complementing raw league tables or a ranking approach with fairer and more accurate approaches, such as value-added, to provide a better picture of school effectiveness.

IELTS Research Program

The IELTS partners, British Council, Cambridge English Language Assessment and IDP: IELTS Australia, have a longstanding commitment to remain at the forefront of developments in English language testing.

The steady evolution of IELTS is in parallel with advances in applied linguistics, language pedagogy, language assessment and technology. This ensures the ongoing validity, reliability, positive impact and practicality of the test. Adherence to these four qualities is supported by two streams of research: internal and external.

Internal research activities are managed by Cambridge English Language Assessment's Research and Validation unit. The Research and Validation unit brings together specialists in testing and assessment, statistical analysis and item-banking, applied linguistics, corpus linguistics, and language learning/pedagogy, and provides rigorous quality assurance for the IELTS test at every stage of development.

External research is conducted by independent researchers via the joint research program, funded by IDP: IELTS Australia and British Council, and supported by Cambridge English Language Assessment.

Call for research proposals

The annual call for research proposals is widely publicised in March, with applications due by 30 June each year. A Joint Research Committee, comprising representatives of the IELTS partners, agrees on research priorities and oversees the allocations of research grants for external research.

Reports are peer reviewed

IELTS Research Reports submitted by external researchers are peer reviewed prior to publication.

All IELTS Research Reports available online

This extensive body of research is available for download from www.ielts.org/researchers.

INTRODUCTION FROM IELTS

This study by Talia Isaacs and her collaborators at the University of Bristol Second Language Speech Laboratory was conducted with support from the IELTS partners (British Council, IDP: IELTS Australia, and Cambridge English Language Assessment) as part of the IELTS joint-funded research program. Research funded by the British Council and IDP: IELTS Australia under this program complements those conducted or commissioned by Cambridge English Language Assessment, and together inform the ongoing validation and improvement of IELTS.

A significant body of research has been produced since the joint-funded research program started in 1995, over 100 empirical studies having received grant funding. After undergoing a process of peer review and revision, many of the studies have been published in academic journals, in several IELTS-focused volumes in the *Studies in Language Testing* series (http://www.cambridgeenglish.org/silt) and in *IELTS Research Reports*. To date, 13 volumes of *IELTS Research Reports* have been produced. But as compiling reports into volumes takes time, individual research reports are now made available on the IELTS website as soon as they are ready.

In the IELTS Speaking test, candidates are assessed according to a number of criteria, pronunciation being one of them. A revision to the way this criterion is assessed was introduced in 2008. Previously, pronunciation was rated on a four-point scale (bands 2, 4, 6 and 8). It was changed to a nine-point scale to bring it in line with the other criteria. In addition, the band descriptors now made examiners consider not just global features of pronunciation, but also specific phonological features that contribute to speech being comprehensible, e.g. chunking, intonation and word stress. Unlike the other criteria, which had descriptors specific to each band level, the descriptors for bands 3, 5 and 7 in pronunciation only say that a candidate "shows all the positive features" of the band below and "some, but not all, of the positive features" of the band above.

Studies conducted with examiners indicate that the revised pronunciation criteria are an improvement, though the evidence also indicates that this criterion remains the most difficult one for them to rate (Galaczi, Lim and Khabbazbashi, 2012; Yates, Zielinski and Pryor, 2011). The current study thus goes one step further and tries to tease out how the various features specified in the band descriptors actually contribute to examiners' scoring decisions.

The results indicate that all the features do contribute to scoring decisions. However, it was also found that no one feature distinguished across bands 5 to 8. Bands 7 and 8, in particular, may not be sufficiently distinguished from one another, (and to a lesser extent, band 5 from band 6). Is this a legacy of the criterion previously having fewer levels? Is it the result of bands 5 and 7 not containing specific performance features of their own? Or is it just that human examiners cannot routinely distinguish that many different levels of pronunciation? It is difficult to tell, and further studies are necessary in this regard.

The study makes clear that, if ever, coming up with a solution that works will be a challenge. The revised pronunciation scale incorporated specific phonological features to help examiners in their decision-making. However, some examiners in this study indicate that considering all those features represent a significant cognitive load, and so might have the opposite effect.

Similarly, multiple descriptors make up each band, and the order in which they are presented may well have an impact. Take band 8 as an example. There is a descriptor asking examiners to consider specific features ("uses a wide range of pronunciation features") and a descriptor asking examiners to make a global judgment ("is easy to understand throughout"), presented in that order. The suggestion is made that simply switching the order in which they are presented would affect the usability of the instrument. The global descriptor helps examiners to quickly determine what band a person is, and they can then use the specific features to confirm that judgment. On the other hand, with this solution, there is a risk that examiners might make the general judgment and not engage with the specifics.

As the foregoing makes apparent, designing mark schemes is not an easy task. The researchers sum it up perfectly: "any revisions to scale descriptors need to find that elusive happy medium between being too specific and too generic and also to take into account considerations of the end-user's cognitive processing when applying the instrument". We could not agree more. Elusive, yes. But IELTS will keep on trying.

Dr Gad S Lim
Principal Research and Validation Manager
Cambridge English Language Assessment

References to the IELTS Introduction

Galaczi, E., Lim, G., and Khabbazbashi, N. (2012). *Descriptor salience and clarity in rating scale development and evaluation*. Paper presented at the *Language Testing Forum*, Bristol, UK, 16-18 November.

Yates, L., Zielinski, E., and Pryor, E. (2011). The assessment of pronunciation and the new IELTS Pronunciation Scale. *IELTS Research Reports*, *12*, pp 23-68.

CONTENTS

1 INTRODUCTION 7

2 LITERATURE REVIEW 7
2.1 Why a focus on the revised IELTS Pronunciation scale? 7
2.2 Previous research on the revised IELTS pronunciation scale 9

3 METHODOLOGY 10
3.1 Research questions 10
3.2 Research design 10
3.3 IELTS speech data 10
3.4 Speaking task and stimulus preparation of audio files for rating 12
3.5 Preliminary study: Piloting the semantic differential scales 12
3.5.1 Background 12
3.5.2 Instrument development, pilot participants, procedure 13
3.5.3 Results of the pilot study 14
3.6 Main study involving IELTS examiners 15
3.6.1 Participants 15
3.6.2 Instruments and data collection procedure 16
3.6.3 Data analysis 17

4 QUANTITATIVE RESULTS 17
4.1 Examiner questionnaire responses: Perceptions of rating linguistic features 17
4.2 Intraclass correlations 18
4.3 Preparation for discriminant analyses 18
4.4 Discriminant analyses 21
4.5 Between-band comparisons for the Speaking and Pronunciation scales 24

5 QUALITATIVE RESULTS 26
5.1 Comparing the retired 4-point with the revised 9-point Pronunciation scale 26
5.2 Assessing pronunciation in relation to other aspects of test-taker ability 26
5.3 Terminology used in the IELTS Pronunciation scale 29
5.3.1 Phonological features and nativeness 29
5.3.2 The in-between IELTS Pronunciation band descriptors 30
5.3.3 Comprehensibility 31

6 DISCUSSION 33
6.1 Summary and discussion of the main findings 33
6.2 Limitations related to the rating instruments and procedure 35

7 REFERENCES 37

APPENDICES

Appendix 1: A description of the 18 researcher-coded measures used in the preliminary study 39

Appendix 2: Background questionnaire 40

Appendix 3: Pre-rating discussion guidelines for focus group 43

Appendix 4: Instructions on rating procedure 44

Appendix 5: Definitions for the constructs operationalised in the semantic differential scales 46

Appendix 6: Instrument for recording ratings for each speech sample 47

Appendix 7: Post-rating summary of impressions 47

Appendix 8: Post-rating discussion guidelines for focus group 48

List of tables

Table 1: Number of test-takers (n = 80) pre-rated at each scale band for the IELTS Speaking and IELTS Pronunciation scale 11
Table 2: Intraclass correlations for the semantic differential scale measures (internal consistency) 14
Table 3: Pearson correlations among the EAP teachers' semantic differential measures for the 40 picture narratives 14
Table 4: Pearson correlations between the discrete semantic differential measures rated by the EAP teachers (n = 10) and the most conceptually similar variables from Isaacs and Trofimovich (2012) 15
Table 5: Means (standard deviations) of IELTS examiners' degree of comfort rating key terms in the IELTS Pronunciation scale (reported as 0 = not comfortable at all, 5 = very comfortable) 18
Table 6: Intraclass correlations for the IELTS examiners' ratings using the IELTS Speaking band descriptors and the semantic differential scales 19
Table 7: Descriptive statistics for target variables used in the discriminant analyses 19
Table 8: Pearson correlations among the Cambridge English pre-rated IELTS Speaking and Pronunciation scores and the UK IELTS examiners' semantic differential ratings 20
Table 9: Summary of global group differences across the four IELTS band placements 20
Table 10: Eigenvalues for discriminant functions 21
Table 11: Structure matrix for IELTS Speaking scores 21
Table 12: Structure matrix for IELTS Pronunciation scores 22
Table 13: Functions at group centroids for IELTS Speaking scores 22
Table 14: Functions at group centroids for IELTS Pronunciation scores 22
Table 15: Classification results for IELTS Speaking scores 24
Table 16: Classification results for IELTS Pronunciation scores 24
Table 17: Summary of univariate ANOVAs for IELTS Speaking scores 25
Table 18: Summary of between-band comparisons for IELTS Speaking bands 25
Table 19: Summary of univariate ANOVAs for IELTS Pronunciation scores 25
Table 20: Summary of between-band comparisons for IELTS Pronunciation bands 26

List of figures

Figure 1: Visual chart showing the mixed methods nature of the research design 11
Figure 2. Discriminant function scores for speaking band placements, with mean centroid values designating IELTS Speaking bands 5 through 8 23
Figure 3: Discriminant function scores for pronunciation band placements, with mean centroid values designating IELTS Pronunciation bands 5 through 8. 23

1 INTRODUCTION

The growing internationalisation of UK campuses has brought with it the concomitant challenge of providing valid assessments of incoming students' English language ability. Higher education institutions often rely on scores from large-scale tests as a measure of prospective students' ability to carry out academic tasks in the medium of instruction for admissions purposes. Due to the high-stakes consequences arising from test score use (both intended and unintended), it is incumbent upon test providers to continue to commit resources to an ongoing and comprehensive program of validating their tests.

One priority area of the IELTS Joint-Funded Research Program in the 'test development and validation issues' category is to examine the 'writing and speaking features that distinguish IELTS proficiency levels' (IELTS, 2014). In light of the 2008 expansion of the IELTS Pronunciation scale from 4- to 9-levels (DeVelle, 2008), there is a pressing need to examine the qualities of test-taker speech that differentiate between Pronunciation scale levels, particularly at the high end of the scale, since these are the levels most relevant for university admissions and, in some cases, international student visa purposes. The present project addresses this gap by examining the linguistic factors that most efficiently distinguish between IELTS Pronunciation levels at the upper end of the scale (IELTS overall band scores of 5 to 8.5). In the next section, we elaborate on our reasons for focusing on the IELTS Pronunciation scale by placing it in the broader context of second language (L2) pronunciation assessment research.

2 LITERATURE REVIEW

2.1 Why a focus on the revised IELTS Pronunciation scale?

Pronunciation is one of the most under-researched areas in language assessment, having been mostly absent from the research agenda since the early 1960s, although there has been a resurgence of interest in pronunciation from within the L2 assessment community against a backdrop of growing momentum among applied linguists and language teachers (Isaacs, 2014). One of the challenges associated with operationalising pronunciation in rating scales is that the theoretical basis for pronunciation in communicatively-oriented models is weak. In Bachman's influential Communicative Language Ability framework (1990) and its refinement in Bachman and Palmer (1996), for example, 'phonology/graphology' appears to be a carryover from the skills-and-components models of the early 1960s (e.g., Lado, 1961). However, the logic of pairing 'phonology' with 'graphology' (i.e., readability of handwriting) is unclear. Similarly, in their model of Communicative Competence, Canale and Swain (1980) do not provide a definition of 'phonology' nor clarify its applicability to L2 learners in particular (as opposed to first language, or L1, learners).

In sum, although developments in language testing and speech sciences research have clearly moved beyond a unitary focus on the applications of contrastive analysis for teaching and testing discrete skills that characterised the skills-and-components models (Bachman, 2000; Piske, MacKay and Flege, 2001), there has been little crossover between these two areas of research. The consequence is that existing theoretical frameworks do not adequately account for the role of pronunciation within the broader construct of communicative competence or communicative language ability.

Because theory often informs rating scale development, it is perhaps unsurprising that pronunciation has not been consistently modeled in L2 oral proficiency scales. In fact, some rating scales exclude pronunciation from rating descriptors (e.g., Common European Framework of Reference benchmark level descriptors; Council of Europe, 2001), which implies that pronunciation is an unimportant part of L2 oral proficiency (Isaacs and Trofimovich, 2012; Levis, 2006). This runs contrary to an increasing consensus among language researchers and teachers and a growing body of evidence that pronunciation is an important part of communication that needs to be addressed through L2 instruction and assessment, particularly in the case of learners who have difficulty being verbally understandable to their interlocutors (Derwing and Munro, 2009; Saito, Trofimovich and Isaacs, 2015).

Pronunciation, and speaking more generally, have had a long history as an assessment criterion in the Cambridge English Language Assessment (hereafter Cambridge English) testing tradition, including in the IELTS test (Weir, Vidaković and Galaczi, 2013). This is in contrast to the Test of English as a Foreign Language (TOEFL), which only included pronunciation as an assessment criterion with the introduction of its speaking component as part of the launch of the internet-based TOEFL (iBT) in 2005 (ETS, 2011). In the context of the Revision Project of the ELTS, which was the direct predecessor test of the IELTS, Alderson (1991) clarified that pronunciation content had not been included in all nine ELTS holistic speaking band descriptors because nine levels might introduce unnecessary or unusable level distinctions for raters. When the IELTS speaking scale was subsequently redeveloped as a 9-point analytic scale, pronunciation was the only one of four subscales to be presented as a 4-point scale and was designated only at even scale levels (2, 4, 6, 8), with no descriptors appearing in the odd bands (1, 3, 5, 7, 9; DeVelle, 2008). However, subsequent research showed that the 4-point scale was too crude in its distinctions (Brown, 2006). More specifically, raters often resorted to band 6 as the 'default' scale levels when rating and were reticent to use band 4, which some expressed was too severe an indictment on the strain incurred in understanding the speech.

This research prompted the expansion of the 4-point Pronunciation scale to a 9-point scale in conformity with the three other IELTS Speaking subscales (DeVelle, 2008). In the wording of the Pronunciation descriptors from the current public version of the scale, which closely resembles the version that accredited IELTS examiners are trained on and use in operational testing settings, Pronunciation scale levels 2, 4, 6, 8, and 9 contain their own unique descriptors (IELTS, 2012). With the exception of Pronunciation scale band 2, in which speech is described as 'often unintelligible' (with no further pronunciation-specific descriptor in band 1 of the public version of the scale), the remaining scale levels 4, 6, 8, and 9 refer to the use of a 'limited range'/ 'a range'/ 'a wide range'/ and 'a full range of pronunciation features' respectively, in the first part of the descriptor for each band, although *which* 'pronunciation features' specifically are being referred to is left undefined (p. 19). In the IELTS examiners' version of the scale, this first part of the descriptor is followed by further specification of selected pronunciation-specific features, including, depending on the band level, rhythm, stress, intonation, articulation of individual words or phonemes, chunking, or connected speech. Finally, by the end of the descriptor, there is some statement about the test-taker's ability to convey meaning or to be understood more or less successfully.

In contrast to these even-level Pronunciation descriptors, Pronunciation scale levels 3, 5, and 7 simply contain the description, 'shows all the positive features of <the scale band immediately below> and some, but not all, of the positive features of <the scale band immediately above>.' The under-specification of pronunciation-specific criteria at these junctures of the scale is unique to the Pronunciation subscale in the IELTS Speaking band descriptors, giving IELTS examiners considerable latitude to assess the test-taker at a level that is in between the specifications of the two levels.

Applicants to UK universities who are required to provide proof of English language proficiency currently need a minimum IELTS score of at least 5.5, equivalent to a Common European Framework of Reference (CEFR) B2 level, in each of the component skills for Tier 4 (student) visa issuance purposes (UK government website, 2014). In practice, research-intensive UK universities tend to require an IELTS Overall Band Score or minimum component scores on each of the subskills of 6.5 or 7.0 to consider an applicant for admission to a program, although there is a degree of variability across universities and departments. The IELTS test is additionally often used as proof of proficiency to gain entry into certain professions or professional programs in the UK and internationally. Following recommendations of a recent standard-setting study conducted in the healthcare sector, for example (Berry, O'Sullivan and Rugea, 2013), the UK General Medical Council recently raised English language proficiency requirements for international doctors wishing to practice in the UK from an IELTS Overall Band Score of 7.0 to 7.5, with each component score necessitating at a minimum of 7.0 (General Medical Council, 2014).

Thus, in such contexts, obtaining a level of 7.0, including on the speaking component, is crucial. However, as described above, the pronunciation component of the scale is not associated with a particular descriptor at band 7, other than that the performance features that the test-taker demonstrates fall between levels 6 and 8 with respect to pronunciation. It follows that in most instances, obtaining an IELTS band 7 is much more consequential for test-takers for gatekeeping purposes (e.g., gaining admission to university or a regulated profession) than obtaining an IELTS band 3 or 5—the other bands for which the pronunciation descriptor suggests that the pronunciation performance is sandwiched between the two adjacent levels. This makes level 7 of particular research interest in the current study, which is set in the UK higher education context.

In light of the latest round of revisions to the Pronunciation component of the IELTS Speaking band descriptors, there is a pressing need to show empirically that, contrary to Alderson's (1991) assertion, raters *can* meaningfully distinguish between nine levels of pronunciation, particularly at the upper end of the scale that is most consequential for high-stakes decision-making in UK universities and beyond. Two recent studies on the revised IELTS Pronunciation scale (Galaczi, Lim and Khabbazbashi, 2012; Yates, Zielinski and Pryor, 2011), which focus on IELTS examiners' self-report data, including their confidence in using the scale and, in the latter study, the pronunciation features they reportedly attend to when scoring, are overviewed in the next section of this report. Although collectively, these studies elucidate examiners' perceptions of discrete scale criteria and perceived difficulty in making level distinctions at different points along the scale, neither study systematically examines the linguistic criteria that are most discriminating at different levels of the IELTS Pronunciation scale—a research gap that the current study seeks to fill.

Yet another reason to investigate the IELTS Pronunciation scale is that there is a need to clarify the underlying construct being measured. The IELTS Speaking scale that accredited IELTS examiners consult in operational testing settings is not currently available for public appraisal. Although a public version of the scale can be accessed in the *IELTS Guide for Teachers* (IELTS, 2012), this guide does not attempt to elucidate the pronunciation construct nor that of any of the other Speaking components, other than to state that the scales are equally weighted to feed into an overall IELTS Speaking band score. In contrast, the 2007 *IELTS Handbook* does provide insight into the notion of the construct being measured, stating that the Pronunciation criterion refers to 'the ability to produce comprehensible speech to fulfil the Speaking test requirements' (IELTS, p. 12). The key indicators of this criterion are further specified as 'the amount of strain caused to the listener, the amount of the speech which is unintelligible and the noticeability of L1 influence'. Munro and Derwing's (1999) conceptually clear definitional distinctions between comprehensibility, intelligibility, and accentedness, which are increasingly pervasive in

L2 pronunciation research (Isaacs and Thomson, 2013), are worthwhile examining here, since these concepts relate to what is described in the IELTS Pronunciation criterion and indicators.

Munro and Derwing (1999) define comprehensibility as listeners' *perceptions* of how easily they understand L2 speech. This construct is operationalised by having raters record their judgments on a rating scale—most often, a bipolar semantic differential scale. Thus, comprehensibility is instrumentally defined, in that it necessitates a rating scale as the measurement apparatus (Borsboom, 2005). Hereafter, the concept of ease of understanding L2 speech will be referred to as 'comprehensibility' when a rating scale is involved, unless the rating scale descriptor or participant's verbatim quotation involves the use of another related term.

In contrast to comprehensibility, intelligibility, or listeners' *actual* understanding of L2 speech, is defined as the amount of speech that listeners are able to understand (Munro and Derwing, 1999). This construct is most often operationalised by calculating the proportion of an L2 speaker's words that the listener demonstrates understanding based on his/her orthographic transcription of an L2 utterance (i.e., percent of words accurately transcribed). From this standpoint, reference to 'comprehensible speech' as the IELTS Pronunciation criterion and to 'listener strain' as the first indicator in the *IELTS Handbook* is consistent with Munro and Derwing's notion of comprehensibility.

Conversely, reference to 'unintelligible' speech and to the 'amount of words' in the second indicator is confusing, since it is listeners' *perceptions* of what they are able to understand that is being captured in the IELTS speaking scale (comprehensibility) and not a word-based understandability count or ratio (intelligibility). These terms are apparently being used interchangeably in the *IELTS Handbook* (IELTS, 2007), but a more nuanced description would be helpful from a research perspective.

Finally, the last indicator, 'the noticeability of L1 influence' evokes the concept of accentedness, defined in the literature as listeners' perceptions of how different the L2 speech sounds from the native-speaker norm (e.g., in terms of discernible L1 features; see Isaacs and Thomson, 2013). Most applied linguists agree that being understandable to one's interlocutor is the appropriate goal for L2 pronunciation instruction (and, by implication, assessment), since L2 learners need not sound like native speakers to successfully integrate into society or to carry out their academic or professional tasks (Isaacs, 2013). Further, L2 speakers with discernible L1 accents may be perfectly understandable to their listeners, whereas speech that is difficult to understand is almost always judged as heavily accented (Derwing and Munro, 2009).

In sum, comprehensibility and accentedness are overlapping yet partially independent dimensions. However, they are often conflated in current L2 oral proficiency scales (Harding, 2013; Isaacs and Trofimovich, 2012), although again, the presence of a detectable accent may have no bearing on a test taker's comprehensibility (Crowther, Trofimovich, Saito and Isaacs, 2014). With regard to the public version of the IELTS Speaking scale, reference to comprehensibility tends to be vague. For example, 'is effortless to understand' or 'mispronunciations are frequent and cause some difficulty for the listener' could benefit from greater precision (IELTS, 2012, p. 19).

In light of the relatively recent expansion of the IELTS Pronunciation scale from four to nine levels, there is a need to bring together different sources of evidence to examine the properties of test-takers' speech (pronunciation) that characterise these different levels of the scale. The next section documents the few recent studies that have been conducted on the IELTS Pronunciation scale specifically, which argues for the need for a more in-depth look at the use of the IELTS Pronunciation scale in relation to pronunciation-specific features.

2.2 Previous research on the revised IELTS pronunciation scale

The current study builds on, complements, and extends previous work on the revised IELTS Pronunciation scale, which, to date, has included two studies. The first consisted of a large-scale worldwide survey conducted within the Research and Validation unit at Cambridge English as part of a larger study (Galaczi et al., 2012). A large sample of accredited IELTS examiners from 68 countries generated 1142 responses about their use of and attitudes toward the IELTS Speaking scale. Results of open- and closed-ended items suggested that examiners understood less of, and were less confident in their use of, the IELTS Pronunciation scale relative to the other three other component Speaking scales. The findings, including examiners' qualitative comments, led the authors to suggest the need for further examiner training with respect to pronunciation to generate clarity around technical concepts (e.g., stress timing, chunking) and elucidate conceptual overlap in terminology (e.g., rhythm, stress, chunking).

Galaczi and her colleagues' (2012) finding about the Pronunciation scale descriptors being more difficult to use relative to descriptors for the other IELTS Speaking subscales was echoed in the first IELTS joint-funded research study to focus on the revised IELTS Pronunciation scale, conducted by Yates and her colleagues (2011). This study involved 27 Australian IELTS examiners first completing a questionnaire on their perceptions of and attitudes toward the revised IELTS Pronunciation scale.

Twenty-six of those examiners then rated 12 IELTS test-takers' speech samples on the IELTS interview task, and those test-takers had been independently rated at each of IELTS Speaking bands 5, 6 and 7. Next, stimulated recalls were elicited from six Australian IELTS examiners who had not participated in the earlier phase of the study. After listening to and scoring the same 12 speech samples, they were asked to pause the recording during a second listening and identify the pronunciation features that had influenced their rating decisions.

Results of descriptive statistics for the questionnaire items and examiners' verbatim comments revealed examiner self-reported difficulty in what one examiner referred to as the 'in between bands,' which referred to bands 5 and 7 in the context of the study (p. 34). Other examiners referred to the vagueness of the descriptors and the recency of the introduction of the pronunciation descriptors leading to greater relative difficulty in conducting assessments using the Pronunciation scale. The authors conveyed examiners' reported difficulty in conducting band level decisions (with adjacent bands naturally proving more difficult to distinguish than non-adjacent bands). They also reported the frequency of the six stimulated recall examiners' comments by pronunciation features, triangulated with the 27 examiners' questionnaire responses of which pronunciation features they deemed most important when conducting their pronunciation ratings. Surprisingly, the authors did not break down reported features that figured into the examiners' decision-making by the test-takers' pre-rated IELTS Speaking levels to reveal the differences in reported features by level. Such an analysis, had it been attempted, would necessarily have been exploratory due to the small sample size of test-takers (four at each level).

To complement and move beyond these findings, which are predominantly based on IELTS examiners' self-report data about their confidence, use of the scale and preferences, there is a need to investigate the trait-relevant criteria that inform these IELTS Pronunciation level distinctions using multiple sources of evidence and to relate these back to the existing Pronunciation descriptors. This is the goal of the present study, with a focus on the levels likely to be most relevant for high-stakes decision-making in UK higher education settings.

3 METHODOLOGY

3.1 Research questions

The current study seeks to identify the linguistic factors that most efficiently distinguish between revised IELTS Pronunciation scale bands. In addition to contributing to the ongoing validation of the IELTS Speaking (Pronunciation) scale, insight into the criteria that raters use to make level distinctions will advance our understanding of the construct of comprehensibility.

The research questions are as follows:

1. *Which speech measures are most strongly associated with IELTS examiners' Pronunciation ratings? Which most effectively distinguish between the upper bands of the IELTS Pronunciation scale?*

2. *How do IELTS examiners engage with the IELTS Pronunciation scale as a component of assessing speaking? What are their perceptions of the rating scale criteria, including the linguistic factors that underlie their Pronunciation scoring decisions?*

Taking into account examiners' perceptions and statistical indices, these findings will be related to the existing IELTS Pronunciation descriptors when interpreting the data, in view of providing recommendations for optimising examiners' use of the scale (e.g., through rater training or scale revisions).

3.2 Research design

The research questions were addressed using a concurrent mixed-methods design (Creswell and Plano-Clark, 2011), with different but complementary sources of data collected during examiner rating and focus group sessions using pre-recorded L2 speech data as stimuli. In the way that the Results section is structured, quantitative analyses are presented first followed by qualitative analyses from the focus group data to bring IELTS examiners' voices to bear in results reporting. A summary of the research design is shown in Figure 1. This visual chart, which breaks down the various phases of the study, can be consulted as a 'roadmap' through the Methodology section that shows the nature of the mixing (see Isaacs, 2013).

3.3 IELTS speech data

Audio recorded speech samples of 80 L2 test-takers (50 female, 30 male) performing the Speaking component of the IELTS were provided by Cambridge English prior to the start of data collection for the current study. The speech samples were collected at 17 test centres around the world, with both the test-taker and the test centres where they were recorded identified using alphanumeric codes in the database to preserve individual and institutional anonymity.

The test-takers were from myriad L1 backgrounds, including Chinese (19), Arabic (16), Tagalog (9), Spanish (6), Thai (5), Kannada (3), and one or two speakers of 14 additional world languages. Table 1 shows the number of test-takers who had been pre-rated at IELTS bands levels 5 to 9, both for the overall Speaking component, and for the Pronunciation subscale. Scores on the other three IELTS Speaking subscales were not provided as part of the dataset, as only the overall IELTS Speaking score is reported to IELTS test users, and this score is the most stable. Access to the Pronunciation subscores for the same test-takers enabled an in-depth investigation of Pronunciation scale band levels in relation to more discrete pronunciation measures in the current study.

Due to the relatively low number of test-takers who had been pre-rated at band 8.5 for Speaking and band 9 for Pronunciation (seven and two test-takers, respectively), these bands were collapsed to form a 'band 8 and higher' category.

A majority of the recorded Speaking performances were reportedly re-marked by multiple IELTS examiners for research purposes using only the audio files as stimuli.

However, a few of the Speaking performances were reportedly scored live during the course of the test (GS Lim, personal communication, April 17, 2013). That is, scoring condition (recorded or live) was not controlled for in the Cambridge English pre-rated data provided for the study nor indicated as a variable in the dataset. Thus, it was unknown to the research team which of the speaking files had been subject to which pre-rated scoring condition.

IELTS scale band	Number of test-takers pre-rated at each level	
	IELTS Speaking scale	IELTS Pronunciation scale
5	23	18
6	19	26
7	23	16
≥8	15	20

Table 1: Number of test-takers (n = 80) pre-rated at each scale band for the IELTS Speaking and IELTS Pronunciation scale

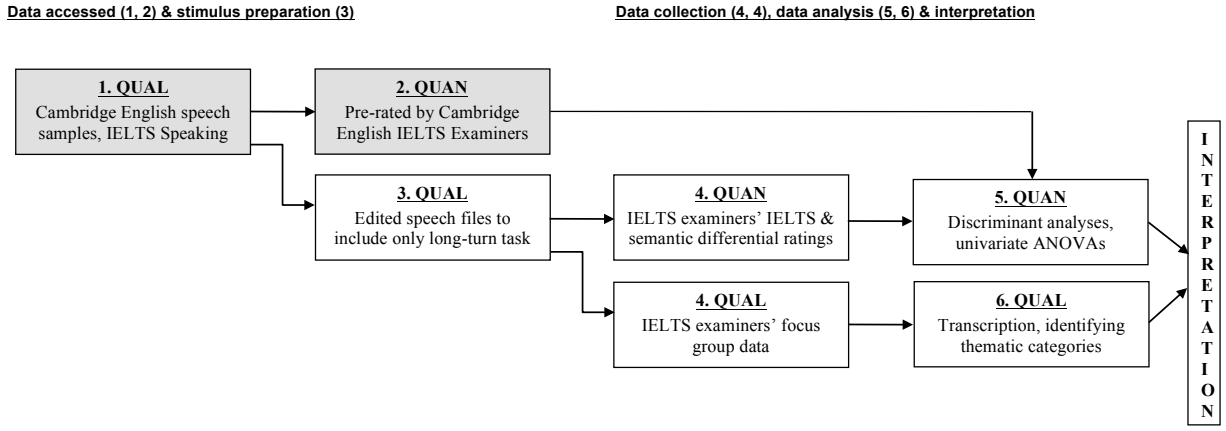


Figure 1: Visual chart showing the mixed methods nature of the research design

Note. Shaded boxes represent IELTS Speaking (1) and pre-rated (2) data provided by Cambridge English prior to the start of the project. Qualitative (QUAL) is used for all non-numerical data and quantitative (QUAN) is used for numerical data only. Because neither QUAL nor QUAN sources of evidence were considered dominant in shedding light on the research phenomenon in this project, CAPS are used throughout. Numbers designate the temporal sequencing of data collection and analysis in carrying out the study. The same numbering for QUAN and QUAL at phase 4 reflects the concurrent nature of data collection, although the results were analysed and reported separately.

3.4 Speaking task and stimulus preparation of audio files for rating

For the purpose of the current study, L2 test-takers' performance on the IELTS long-turn speaking task (task 2) was used for rating. Although all three IELTS speaking tasks had figured into the Cambridge English pre-rated scoring, it was not feasible to include test-takers' entire speaking performance within the confines of the study. One reason for the selection of the long-turn task was that a more monologic task that minimises variability in interviewer style as part of the performance (Brown, 2005) would likely promote greater rater (IELTS examiner) focus on the quality of the test-taker's language rather than on his/her exchanges with the interviewer. A second reason is that the majority of current L2 pronunciation studies are conducted using monologic tasks (Isaacs and Thomson, 2013), which would bring this study in line with that body of second language acquisition (SLA) oriented pronunciation research. Further, the intention was to analyse L2 speech data using measures that mostly stem from a cognitive (psycholinguistic) view of language, in accordance with previous research on the linguistic factors that underlie the 'comprehensibility' construct (Isaacs and Trofimovich, 2012; Saito et al., 2015). These measures, discussed below, would have needed to have been adapted considerably to accommodate the complexities of interactional data (e.g., turn-taking, floor-holding strategies; Ejzenberg, 2000), making the long-turn task, with its attempt to elicit sustained speech, the best option for further analysis.

To prepare the spoken stimuli for rating, each test-taker's long-turn task performance was excised from the recording immediately after the interviewer's initial prompt until the conclusion of the task ($M_{duration}$= 128 seconds; 59–232 seconds). The audio data were of highly variable sound quality, having been recorded at 17 different IELTS test centres. While some files were of reasonable sound quality, others were extremely poor, to the extent that it was difficult to discern what was being said. Some of the recording problems included the buzz or hiss of the recording device drowning out the speech, inadequate recording volume, or the impromptu incursion of distracting background noise at various junctures throughout the performance (e.g., sirens). The accredited IELTS examiners who conducted the Cambridge English pre-ratings were apparently able to score the speech despite these recording quality difficulties. On this basis, no files with poor recording quality were discarded nor was editing individual sound files feasible, since this treatment would not have been uniform across files that had already been pre-rated. Instead, the entire batch of audio files was edited to optimise the sound quality using Adobe Audition C36 version 5.32 and WavePad Sound Editor version 5.33.

The editing steps applied to the batch of the 80 files included:

1. converted all files to mono channel
2. normalised the files to 85% peak intensity
3. applied DC offset correction to centre soundwaves (correct skew)
4. applied noise reduction (auto spectral subtraction; silence to audio proportion: 30%)
5. applied dynamic range compressor at the general voice level preset to ensure that sample volume stays within a prescribed range (threshold -20dB, ratio: 4:1, limit: 0dB) to correct clipping due to input (microphone) levels being too high during recording

Even after applying these procedures to all files, the sound quality of a portion of the files remained poor, representing a confound for a study examining the construct of comprehensibility (i.e., not clear if it is the speech itself or the poor audio quality that results in perceived strain on the part of the listener in terms of understanding the message). The variable quality of the L2 speech files also proved prohibitive for undertaking analyses of the data using auditory and instrumental measures in line with previous L2 pronunciation research (Isaacs and Trofimovich, 2012; Trofimovich and Isaacs, 2012) as was the original plan for the project. In order to move beyond this limitation, a preliminary study was conducted to pilot a new procedure for obtaining discrete listener-rated measures of pronunciation and of other linguistic features using semantic differential scales. The validation of this procedure is described in the next section, using previous research as the starting point.

3.5 Preliminary study: Piloting the semantic differential scales

3.5.1 Background

Recent studies by Isaacs and Trofimovich (2012) and Trofimovich and Isaacs (2012) aiming to 'disentangle' accent from comprehensibility, are foundational to the current study. The approach was to elicit 60 native English listeners' L2 accentedness and comprehensibility ratings based on English picture narratives spoken by 40 adult L1 French learners in the Canadian context. The listeners' mean accentedness and comprehensibility ratings, obtained using 9-point Likert-type scales used by convention in L2 pronunciation research (Isaacs and Thomson, 2013), were then correlated with 18 researcher-coded measures derived from the speech samples, including both instrumental measures (obtained using speech analysis software), and auditory measures. These measures spanned the domains of pronunciation, fluency, lexicogrammar, and discourse. Appendix 1 describes how each measure was computed, and examples from L2 learner data can be found in the original articles.

By bringing together results of statistical analyses and experienced L2 teacher-raters' perspectives on the linguistic influences on their judgements from introspective reports, a subset of measures that best distinguished between three levels of L2 comprehensibility for L1 French leaners of English were identified. Lexical richness and fluency measures distinguished between low levels of comprehensibility, grammatical and discourse-level measures distinguished between high levels, and word stress distinguished between all three comprehensibility levels examined.

In terms of 'disentangling' accent from comprehensibility, the major finding was that accentedness is principally linked to pronunciation-specific linguistic features, including rhythm and segmental (i.e., vowel and consonant) accuracy. Conversely, comprehensibility cuts across a much wider range of linguistic variables than simply pronunciation, with lexical richness and grammatical accuracy also contributing to the variance in comprehensibility ratings along with word stress. Further research examining L1 effects has demonstrated the robustness of the finding that accentedness relates chiefly to linguistic variables subsumed under the umbrella term 'pronunciation' including segmental and prosodic (i.e., stress, rhythm, intonation) variables, whereas comprehensibility is linked to both pronunciation (e.g., segmental errors, word stress, intonation, speech rate) and lexicogrammatical dimensions (lexical richness and appropriateness, grammatical accuracy and complexity, discourse measures; Crowther et al., 2014; Saito et al., 2015).

3.5.2 Instrument development, pilot participants, procedure

As referred to earlier in the report, the original intention was to adopt Isaacs and Trofimovich's (2012) and Trofimovich and Isaacs' (2012) methodology to obtain auditory and instrumental measures derived from each test-taker's performance on the IELTS long-turn task. The novel aspect would be relating these measures to IELTS scores rated by accredited IELTS examiners using the IELTS Speaking band descriptors (as opposed L2 comprehensibility and accentedness ratings scored by lay listeners on Likert-type scales in the context of those published studies, which was far removed from a high-stakes assessment context). However, several recorded passages proved untranscribable into standard orthography due to poor recording quality, making it impossible to obtain the auditory and instrumental measures as planned.

As a result of this logistical challenge, the alternative procedure of developing semantic differential scales with which to record the IELTS examiners' ratings of both comprehensibility, and more discrete linguistic measures of L2 speech was proposed, trialled, and ultimately implemented. In order to examine the efficacy and pilot the methodology of using semantic differential scales to capture IELTS examiners' discrete ratings of linguistic features as an alternative to the more objective researcher-coded auditory and instrumental measures reported in Isaacs and Trofimovich (2012) and Trofimovich and Isaacs (2012), it was desirable to trial the use of those scales in the original context of those studies using the same 40 L1 French speech samples. The semantic differential measures obtained as a result of piloting could then be related to the more objective original measures generated in that study. To this end, ratings of 10 experienced English Canadian-born English for Academic Purposes (EAP) teachers who reported having normal hearing were elicited to provide baseline data for the main UK-based IELTS project described below. The Canadian EAP teachers reported speaking English on average 93% of the time daily (SD = 8.2) and estimated having 11.7 years of ESL teaching experience (SD = 8.6), including 7.9 years of EAP-specific experience (SD = 7.6). Seven out of the 10 teachers reported having received university-level pronunciation training (e.g., phonology for teachers).

Printed copies of the semantic-differential scales were constructed using 5 cm lines and separate endpoint descriptors for each scale, with a frowning face at the leftmost (negative) end and a smiley face at the rightmost (positive) end of the spectrum. No marked intervals nor numerical endpoints were indicated on the scale. The EAP teachers were instructed to mark an 'X' on each scale (line) to record their ratings, and their score was later computed by measuring the placement of the 'X' manually with a ruler. The teachers performed the semantic differential scale ratings in a fixed order, starting with a global rating of L2 comprehensibility. This was measured on a continuum ranging from 'painstakingly effortful to understand' to 'effortless to understand' following a study by Isaacs, Foote and Trofimovich (2013), which had established, through a consultation with teacher-raters, that this was a clear-cut and user-friendly description of the polar extremes of L2 comprehensibility that conformed with the psycholinguistic aspect of the degree of perceived listener processing effort in understanding L2 speech. This initial listening of the speech sample for a given L2 speaker was immediately followed by eliciting more discrete ratings of seven linguistic variables using separate semantic differential scales during a second listening. The measures included vowel and consonant errors, word stress, intonation, speech chunking, speech rate, lexical richness, and a combined measure of grammatical accuracy and sentence structure. This last measure was grouped together in one scale so as not to exceed seven scales that the raters needed to complete during the second listening.

The wording of the semantic differential scales was selected to roughly correspond to terminology that appeared in the examiners' version of IELTS Pronunciation scale (e.g., chunking), with an attempt to incorporate into the instrument terms which the IELTS examiners, who would take part in the main study, would be familiar with from the scale (see below). Because the examiners' version of the IELTS Speaking band descriptors is not currently available in the public domain and none of the Canadian EAP teachers were IELTS

examiners and, hence, were not privy to the examiners' version of the scale, care was taken to ensure that the wording of the scalar endpoints and accompanying definitions developed for each semantic differential scale did not too closely resemble the wording in the IELTS examiners' version of the scale for intellectual property reasons. In fact, the EAP teachers were not informed about the relation of the pilot study to the IELTS scale. Because the precise definitions of the terms used in the IELTS Speaking band descriptors are often unclear or unspecified in the *IELTS Handbook* (IELTS, 2007), descriptors of the seven discrete measures were drafted based on standard uses of the terms in the literature. Although measures of speech rate, lexical richness and grammatical accuracy/sentence structure are beyond the remit of the IELTS Pronunciation scale, they were included as semantic differential measures due to findings from previous studies about their role in underlying listeners' L2 comprehensibility ratings (Saito et al., 2015; Trofimovich and Isaacs, 2012). In order for the teacher raters to achieve a baseline understanding of the meaning of the terms, the teachers were provided with the definitions shown in Appendix 5 to accompany the semantic differential scalar endpoint descriptors shown in the second part of Appendix 4.

3.5.3 Results of the pilot study

Table 2 shows intraclass correlation coefficients calculated from the 10 EAP teachers' judgments using the semantic differential scales, revealing high internal consistency (.91-.96) for each of the rated measures. Notably, the measures that yielded the highest coefficients and, thereby, the greatest rater consensus tended to either be global measures of the L2 speech (comprehensibility) or non-pronunciation-measures (lexical richness, grammar/sentence structure, and speech rate). Vowel and consonant errors, word stress, and speech chunking yielded slightly lower coefficients, signalling that raters had greater difficulty achieving consensus for these more discrete measures, albeit by a small margin. The scale that yielded the lowest intraclass correlation coefficient and smallest degree of rater agreement was intonation. This construct could arguably be considered more inherently elusive than the other pronunciation measures, as intonation is among the most difficult aspects of pronunciation to teach (Setter, 2005) and is not notated in the English orthographic system. Thus, noticing intonation errors may be less clear-cut than some of the other features. For the purpose of this study, all intraclass correlations were deemed sufficiently high to proceed with further analyses (> .9).

EAP teacher-rated semantic differential measures	Intraclass correlations
Comprehensibility	.95
Vowel & consonant errors	.93
Word stress	.93
Intonation	.91
Speech chunking	.93
Speech rate	.94
Lexical richness	.96
Grammatical accuracy & sentence structures	.95

Table 2: Intraclass correlations for the semantic differential scale measures (internal consistency)

Pearson correlations among the 18 researcher-coded auditory and instrumental measures reported in Isaacs and Trofimovich (2012) and the 9-point scalar comprehensibility rating, pooled over the 60 raters in that study, yielded coefficients ranging from an absolute value of .32 and .78, with six measures above |.7| (tokens, word stress errors, vowel reduction errors, mean length of run, and story breadth). Table 3 shows the Pearson correlations among the semantic differential measures. Due to the fact that these indices were derived using the same semantic differential scale format of marking L2 learners' ability on separate 5 cm lines (as opposed to the researcher-coded measures from Isaacs and Trofimovich, which used different metrics), it is perhaps unsurprising that the semantic-differential measures were strongly correlated.

Measures	1	2	3	4	5	6	7
1 Comprehensibility							
2 Vowel & consonant errors	.905						
3 Word stress	.905	.934					
4 Intonation	.900	.901	.954				
5 Speech chunking	.944	.920	.953	.944			
6 Speech rate	.909	.808	.868	.872	.936		
7 Lexical richness	.938	.913	.882	.891	.930	.860	
8 Grammatical accuracy	.940	.925	.896	.905	.931	.835	.966

Note. All correlations significant at the $p < .01$ level

Table 3: Pearson correlations among the EAP teachers' semantic differential measures for the 40 picture narratives

Semantic differential measures, 7 discrete scales	Conceptually related researcher-coded measures	*r*
Vowel & consonant errors	Segmental error ratio	-.640
Word stress	Word stress error ratio	-.695
Intonation	Rhythm Pitch contour	.715 .535
Speech chunking	Types Mean length of run Rhythm	.828 .782 .770
Speech rate	Mean length of run	.748
Lexical richness	Types	.881
Grammatical accuracy & sentence structure	Grammatical error ratio	-.644

Note. All correlations significant at the $p < .01$ level.

Table 4: Pearson correlations between the discrete semantic differential measures rated by the EAP teachers (n = 10) and the most conceptually similar variables from Isaacs and Trofimovich (2012)

Table 4 shows correlations between the EAP teachers' semantic differential scale ratings (Appendices 5 and 6) and the researcher-coded measures from Isaacs and Trofimovich (2012) that were considered to be most conceptually similar (Appendix 1; see also Saito, Trofimovich, and Isaacs, in press). In the case of some semantic differential measures, more than one conceptually related variable is reported if they are deemed to encompass more than one dimension. For example, literature on intonation has emphasised both a rhythm/timing and a pitch/melody component that operate in tandem (Chun, 2002). In the absence of a specific definition for 'chunking' in the *IELTS Handbook* (2007) and *IELTS Guide for Teachers* (2012), the definition provided to raters in Appendix 5 was:

> *When speaking, people naturally break speech into chunks. For example, when someone says 'how are you,' it is said as one smooth chunk without any pausing. If pauses come in unnatural places, then there are problems with speech chunking.*

As reflected in the definition, we interpreted 'speech chunking' to incorporate the notion of appropriate speech rate and pausing at logical junctures and to imply the use of collocations or memorised chunks facilitated by automatised lexical retrieval (Segalowitz, 2010). A combination of researcher-coded measures related to lexical choice, temporal fluency, and rhythm/timing was interpreted as consistent with this definition and is reported in Table 3. Correlations with types is reported in lieu of tokens due to extremely high intercorrelations between these lexical variables in the Isaacs and Trofimovich (2012) study ($r > .95$).

In all cases, moderate to strong correlations between conceptually related semantic differential and researcher-coded measures were obtained. This result, coupled with high Cronbach's alpha coefficients derived using the semantic differential scales, provided an empirical baseline and justification for the using these scales with UK-based IELTS examiners as a means of understanding the linguistic criteria that best distinguish between upper levels of the revised IELTS Pronunciation scale in the current study.

3.6 Main study involving IELTS examiners

3.6.1 Participants

Eight accredited IELTS examiners (6 female, 2 male) born in the UK (6), Australia (1), and Belgium (1; English-dominant early bilingual), all of whom had resided in the UK for decades and who were affiliated with an IELTS test centre at a research-intensive university in Southwest England at the time of data collection, participated in the study (M_{Age} = 52.5 years; range: 37–66). For the purposes of data entry and reporting, individual examiners were given the pseudonyms E1–E8 to safeguard their anonymity. The examiners functioned predominantly in English in their daily lives and reported speaking English 100% (6), 90% (1), or 80% (1) of the time. They were highly experienced ESL/EFL professionals, with an average of 19.5 years of L2 teaching experience (range: 12–31). All were holders of Certificate in Teaching English to Speakers of Other Languages (CELTA) and/or Diploma in Teaching English to Speakers of Other Languages (DELTA) teaching qualifications, and three additionally held Master's degrees in applied linguistics or linguistics.

All participants reported working as IELTS examiners on average for 8.4 years (range: 2–13). In conformity with the regulations, all had completed their last IELTS recertification and associated quality control monitoring less than two years prior to the start of data collection, which was four or five years after the revised IELTS pronunciation scale had first been introduced.

Two examiners reported having a background in phonetics as part of their teacher training and two others reported having taken a workshop centring on the use of the English phonemic chart as part of British Council training. The remaining examiners reportedly had had no background in pronunciation, but, as with all examiners, had exposure to pronunciation through their use of the IELTS speaking band descriptors, which they routinely used to derive Pronunciation subscores. Due to the small local pool of accredited IELTS examiners, all of whom needed to sign confidentiality agreements and be authorised by the British Council to participate in the study, piloting was bypassed so that all eligible volunteers could take part in the main study.

3.6.2 Instruments and data collection procedure

Data collection consisted of three sessions conducted on three separate days capped at 12 hours total (i.e., four hours/day), although most examiners completed the tasks in 10 hours and all completed data collection within an eight-day span. All sessions consisted of focus groups of two or three examiners, with the composition of the groups varying day by day due to the examiners' scheduling availability. To mitigate examiner (rater) fatigue, planned breaks took place after focus group debriefs. In addition, during the self-paced individually-conducted ratings, examiners were at liberty and encouraged to take short, frequent breaks as they needed.

On the first day of data collection, following the completion of written consent forms, the examiners completed the background questionnaire shown in Appendix 2. This instrument was adapted from a questionnaire developed for UK EAP teachers from an earlier study (Isaacs et al., 2013), with IELTS-specific questions added. Examiners were probed about their language and assessment background in addition to their comfort in making level distinctions for Pronunciation relative to the other Speaking component scales and with the terminology used in the IELTS Pronunciation scale. Semantic differential scales were used to capture their responses, with smiling or frowning face used to signal positive and negative response extremes. Responses were measured by a research assistant with a ruler following data collection (0 = not comfortable at all; 5 = very comfortable).

Next, examiners participated in pre-rating focus group reflections, with the researcher's semi-structured prompts mainly designed to gauge their impressions of the IELTS Pronunciation scale descriptors, level distinctions, interpretation of constructs operationalised in the scale, and impressions of examiner training (see Appendix 3). Thereafter, for the main rating session, each examiner was asked to listen to the L2 speech samples presented in a unique randomised order via headsets connected to the examiner's designated laptop and to independently rate test-takers' IELTS long-turn task performances using printed scoring booklets (Appendix 6).

Hard copies of the rating instructions (Appendix 4), official IELTS Speaking band scale descriptors (confidential), and definitions to go along with the semantic differential scales (Appendix 5) were also provided to examiners to consult during the rating session.

Prior to conducting the ratings IELTS examiners were orally briefed on the possible set of speaking topics that would arise in the set of ratings on the long-turn task, as they would not hear the IELTS interviewer's scripted prompt in the edited audio files. Of nine possible topics, the three most frequent in the dataset were about travel plans (28), neighbours (15), and aspirations (14). The examiners were also informed about the variable sound quality of the speech files prior to starting rating, as per the written description in Appendix 4. They were told that if any of the particular rated measures were, in their view, impossible to reasonably assess due to the poor recording quality, they could leave this rating blank and indicate that on their rating response sheet but were directed to do this only as a last resort. Taken together, less than 2% of the data were deemed unassessable by four of the eight IELTS examiners, representing a small proportion of the data. However, the compromised sound quality and other artifacts of the research setting (e.g., scoring pre-recorded speech vs. live performances; basing ratings on only one vs. on all three IELTS Speaking tasks) inevitably arose in the focus group discussions as limitations of the research setting.

After reading over the rating instructions in Appendix 4, which explained the rating procedure, the IELTS examiners conducted their first listening while consulting the IELTS Speaking band descriptors and scoring the performance using the scale on the left hand side of the page in the rating booklet (Appendix 6). During a second listening, they scored the speech for all eight semantic differential scale measures described in the preliminary study, which included comprehensibility, vowel-consonant errors, word stress, intonation, speech chunking, speech rate, lexical richness, and grammatical accuracy-sentence structure. Although the examiners were instructed to listen to the entire audio file (task performance) in their first listening before conducting their ratings on the IELTS Speaking scale, it was at the examiners' discretion as to how much of the audio file they needed to listen to again to complete the eight semantic differential ratings. They recorded their ratings using the 5 cm semantic differential scales provided on the right hand side of the scoring booklet (Appendix 6). As in the preliminary study, the location of the 'X' mark on all semantic differential scales was manually calculated by a research assistant following data collection using a ruler.

Due to the large volume of speech samples to be assessed (80 test-takers, 2 hrs 51 min recordings time), most of the time during the sessions was spent conducting independent ratings, with a quick debrief about any issues encountered at the end of each 4 hour session.

Finally, in the final session on the third day, the examiners were guided in reflecting on the linguistic criteria that were most salient in informing their judgments. The post-rating questionnaire in Appendix 7 asked examiners to summarise their impressions of which linguistic aspects of speech had contributed most to their impressions of ease/difficulty of understanding the speech (i.e., comprehensibility) and to underscore any additional criteria that could be useful in rating Pronunciation using the IELTS scale. This was used as a springboard for the final focus group discussion using the post-rating guiding questions shown in Appendix 8, mostly recasting some of the earlier issues that had been raised (e.g., descriptor interpretation, scale level distinctions) and probing any other impressions or issues that arose during their data collection experience.

3.6.3 Data analysis

Following the presentation of descriptive statistics (examiners' impressions of rating from the background questionnaire, the internal consistency of their ratings, etc.), the crux of the quantitative analysis is discriminant analyses to examine the clustering and discriminability of the semantic differential measures at the upper levels of the IELTS pre-rated Speaking and Pronunciation scales. These analyses followed by univariate ANOVAs were run to examine the patterning of the linguistic measures and provide an empirical basis for examining scale band distinctions.

In terms of qualitative analyses, over 6 hours (> 50,000 words) of audio recorded focus group data were orthographically transcribed by a research assistant who was present for most of the focus group sessions. The analytic approach was to retain examiners' verbatim comments but to group key quotes or multi-turn focus group exchanges together under provisional headings linked to the research aims and questions. These were then formalised under thematic categories in the Results section, also expressed as headings, and interspersed with the researchers' commentary and interpretation. In accordance with the research aims and questions, data reporting will focus primarily on comments that elucidate raters' perceptions of the IELTS Pronunciation scale, including their understanding of key terminology and description of the processes involved in arriving at scores at critical junctures. The qualitative data will be used to extend and explain the statistical findings and crucially to give voice to examiners as key stakeholders and end-users of the scale.

4 QUANTITATIVE RESULTS

4.1 Examiner questionnaire responses: Perceptions of rating linguistic features

In the background questionnaire prior to listening to speech samples and conducting speech ratings as part of the study, IELTS examiners were asked to compare the IELTS Speaking components for degree of comfort in providing ratings based on their IELTS examining experience. The examiners were most uniformly comfortable making level distinctions using the Lexical Resource subscale (M = 4.55; SD = .07) followed by Grammatical Range and Accuracy (M =4.12; SD = .33) and Fluency and Coherence (M = 3.93, SD = .95). They were the least comfortable providing IELTS Pronunciation ratings (M = 3.29; SD = .86), which is consistent with both the Galaczi et al. (2012) and Yates et al. (2011) studies, providing justification for probing pronunciation further in the current study.

Raters were also asked about their comfort using the terminology featured in the IELTS Pronunciation scale (examiner's version) in the background questionnaire based on their rating experience. Table 5 shows that examiners were nearly uniformly comfortable with 'intelligibility' (the term referred to in the scale, but with the same meaning as narrowly-defined 'comprehensibility' in the current study). In contrast, terms such as phonemes (i.e., individual vowel and consonant sounds) and stress-timing engendered less comfort. Notably, 'phonological features' received by far the lowest rating and was the most variable of all rated terms. As mentioned in the Literature review, specification of the phonological features being referred to is not elaborated on in the scale descriptors and, thus, is subject to considerable interpretative leeway by IELTS examiners.

In the final questionnaire at the end of the focus group sessions (Appendix 7), which asked examiners to rank order the linguistic factors, as represented in the semantic differential scales, that they perceived to have most influenced their understanding of the speech, three of the eight examiners revealed that speech rate had been most influential. Opinion was divided on the rest of the pronunciation-relevant choices, with one examiner selecting each of word stress errors, vowel and consonant errors, and intonation. No examiners ranked either lexical choice or grammatical accuracy and sentence structure as being most important for comprehensibility—variables that had been included as rated features in the semantic differential scales due to their empirical association with comprehensibility in previous L2 pronunciation research (e.g., Crowther et al., 2014; Isaacs and Trofimovich, 2012) but that were beyond the scope of 'Pronunciation' in the context of the IELTS Speaking band descriptors.

In sum, there was no consensus on the most important pronunciation-relevant features for examiners' understanding of the speech, other than that speech rate appeared to be a factor. However, one examiner ranked speech rate as being the least consequential of all pronunciation features for comprehensibility. Such idiosyncratic perceptions are consistent with previous L2 pronunciation research that has highlighted rater difficulty in pinpointing and agreeing on the precise cause of communication breakdowns (Isaacs, 2008; Isaacs and Trofmovich, 2012).

Measure	*M*	*SD*
Intelligibility	4.5	.14
Stress	4.05	.64
Speech rate	4.00	.83
Connected speech	3.95	.64
Accent	3.80	.99
Rhythm	3.78	.71
Intonation	3.70	1.13
Chunking	3.47	1.07
Phonemes	3.46	1.31
Stress-timing	3.36	1.22
Phonological features	2.82	1.45

Table 5: Means (standard deviations) of IELTS examiners' degree of comfort rating key terms in the IELTS Pronunciation scale (reported as 0 = not comfortable at all, 5 = very comfortable)

4.2 Intraclass correlations

Intraclass correlation coefficients, shown in Table 6, were computed to examine the internal consistency of the eight IELTS examiners' ratings for both the IELTS Speaking scales (component and overall scores), and the semantic differential scales in the context of the research study based on their completed ratings. Predictably, IELTS examiners achieved greater overall consensus in assigning scores using the IELTS scales (range: . 838-.867) than on the semantic differential scales (range: .539-795), for which they had received no training and which lacked level demarcations besides the scalar endpoints. The fact that the intraclass correlation was lower for Pronunciation than for the other IELTS subscales is consistent with the IELTS examiners' reported lower comfort level assessing Pronunciation relative to the IELTS Speaking subscales in their background questionnaire and conforms with raters' perspectives from previous studies (Galaczi et al., 2012; Yates et al., 2011).

In terms of intraclass correlations for the semantic differential scale measures, the same overall pattern is revealed for the UK IELTS examiners' scores as for the 10 Canadian EAP teachers' ratings. Coefficients are highest for non-pronunciation related constructs (lexical richness, grammatical accuracy and sentence structure), signalling the greatest examiner agreement for those measures, whereas intonation and vowel and consonant errors are the least internally consistent. Although intraclass correlations obtained from the UK IELTS examiners were considerably lower than the semantic differential scales elicited from the Canadian EAP teacher raters in the preliminary study (.91-.96), with only the former group rating participants from different L1 backgrounds, coefficients were still considered high enough for research purposes to proceed with further statistical analyses.

4.3 Preparation for discriminant analyses

Discriminant analyses were carried out to investigate the contributions of the eight semantic differential speech measures (comprehensibility, segmental errors, word stress, intonation, speech chunking, speech rate, lexical richness and grammatical accuracy) to explain speakers' level placement across several bands of the IELTS Speaking and Pronunciation scales (Band 5, Band 6, Band 7, and Band 8 and over), as pre-rated in the data provided by Cambridge English. Discriminant analyses are ideally suited for this purpose because they allow researchers to separate (discriminate) several data groups (in this case, IELTS band placements) using multiple predictor variables (in this case, eight speech measures), thus minimising the possibility of misclassifying cases into their respective groups or categories. In this section, descriptive statistics and results of diagnostic tests checking the assumptions of discriminant analyses are presented. These are precursors to reporting the results of two discriminant analyses using IELTS Speaking and Pronunciation band placements as grouping variables in the next section. Finally, follow-up univariate ANOVAs examining between-group differences in band placement for each speech measure are calculated separately.

Table 7 shows descriptive statistics for all target variables used in the analyses. As shown in this table, the IELTS Speaking and Pronunciation scores were similar, both in terms of measures of central tendency and variability. The eight speech variables differed slightly, with a number of missing data points for several measures (e.g., 10 for speech chunking and speech rate, 17 for segmental errors) and with means ranging between 28.0 (speech chunking) and 31.9 (comprehensibility).

Scales	Speech measures	Intraclass correlations
IELTS Speaking overall and component scales	IELTS Fluency & Coherence	.845
	IELTS Lexical Resource	.867
	IELTS Grammatical Range & Accuracy	.843
	IELTS Pronunciation	.838
	IELTS Speaking overall	.860
Semantic differential scales	Comprehensibility	.725
	Vowel & consonant errors	.663
	Word stress	.743
	Intonation	.539
	Speech chunking	.731
	Speech rate	.730
	Lexical richness	.795
	Grammatical accuracy & sentence structure	.792

Table 6: Intraclass correlations for the IELTS examiners' ratings using the IELTS Speaking band descriptors and the semantic differential scales

Measure	*n ratings*	*M*	*SD*
IELTS Speaking score[1]	640	6.3	1.0
IELTS Pronunciation score[1]	640	6.4	1.0
Comprehensibility[2]	624	31.9	10.7
Vowel & consonant errors[2]	623	28.8	11.1
Word stress[2]	624	30.0	10.3
Intonation[2]	629	28.8	16.7
Speech chunking[2]	630	28.0	11.2
Speech rate[2]	630	28.5	11.0
Lexical richness[2]	627	28.8	10.2
Grammatical accuracy & sentence structure[2]	628	29.0	10.0

Note. Measured on [1] the 9-point IELTS Speaking band descriptors [2] the semantic differential scales and reported out of 50

Table 7: Descriptive statistics for target variables used in the discriminant analyses

Intercorrelations between the Cambridge English pre-rated IELTS Speaking and Pronunciation scores and the semantic differential measures assigned by the IELTS examiners are shown in Table 8. Notably, comprehensibility correlated the least strongly with both the IELTS Speaking and IELTS Pronunciation scores of all semantic differential measures examined (.509 and .476, respectively). However, comprehensibility was strongly correlated ($r > .70$) with the more discrete measures of vowel and consonant (i.e., segmental) errors, word stress, speech rate and speech chunking variables. The relatively lower correlations with grammatical accuracy and lexical richness are plausible in light of recent research suggesting that the variance in rater-assigned L2 comprehensibility scores is differentially explained by pronunciation and lexicogrammatical dimensions depending on the learners' L1 background. In Crowther et al. (2014), for example, in the case of L1 Chinese learners of English, who were the dominant L1 in the present study ($n = 19$), pronunciation variables, and particularly segmental errors, had bearing on their comprehensibility scores.

By contrast, for Hindi-Urdu learners, of which there were four in the present study, lexicogrammatical dimensions of lexical richness and appropriateness, discourse richness, and grammatical complexity contributed to L2 comprehensibility ratings; however, pronunciation-related variables did not affect their comprehensibility scores.

To determine whether it was worthwhile proceeding with the discriminant analyses, it was first necessary to examine whether there was a significant difference between the IELTS bands for the speech variables (semantic differential scale measures). Results of Wilks' lambda tests in Table 9 show that all eight speech measures differed across group levels, implying large separation between groups based on all speech variables. Therefore, the assumption that the IELTS band placements differ as a function of the eight speech measures was supported.

Measure	1	2	3	4	5	6	7	8	9
1 IELTS Speaking score									
2 IELTS Pronunciation score	.914								
3 Comprehensibility	.509	.476							
4 Segmental errors	.585	.578	.885						
5 Word stress	.710	.666	.870	.910					
6 Intonation	.629	.633	.649	.624	.733				
7 Speech chunking	.766	.717	.760	.749	.856	.819			
8 Speech rate	.712	.665	.814	.749	.849	.679	.869		
9 Lexical richness	.818	.780	.646	.713	.832	.680	.841	.829	
10 Grammatical accuracy	.821	.791	.670	.746	.840	.702	.855	.840	.978

Note. All correlations significant at the $p < .01$ level.

Table 8: Pearson correlations among the Cambridge English pre-rated IELTS Speaking and Pronunciation scores and the UK IELTS examiners' semantic differential ratings

Measure	*Wilks' lambda*	*F*-value	*df*	*p*
Comprehensibility	.919	17.749	3603	.0001
Segmental errors	.902	21.918	3603	.0001
Word stress	.854	34.338	3603	.0001
Intonation	.822	43.463	3603	.0001
Speech chunking	.820	43.983	3603	.0001
Speech rate	.852	34.977	3603	.0001
Lexical richness	.781	56.280	3603	.0001
Grammatical accuracy & sentence structure	.755	65.306	3603	.0001

Table 9: Summary of global group differences across the four IELTS band placements

We then tested the assumption of the equality of variances and covariances across all four IELTS level placement bands. However, Box's test exploring this assumption yielded a significant value, $F(108, 673922.57) = 4.32$, $p < .0001$, suggesting that variances and covariances across the data in the four IELTS placement bands were unequal and that the assumption of homogeneity was violated. Based on these results, discriminant analyses should be interpreted conservatively. The final preliminary analysis involved running two separate MANOVAs, with IELTS placement bands for Speaking and Pronunciation used as grouping variables and the eight speech variables used as dependent variables. The two MANOVAs, which are considered standard tests preceding discriminant analyses (e.g., Field, 2009), yielded significant effects of group (using Pillai's trace) for both IELTS Speaking band placement, $V = .30$, $F(24, 1794) = 8.38$, $p < .0001$, and IELTS Pronunciation band placement, $V = .28$, $F(24, 1794) = 7.79$, $p < .0001$. Because Pillai's criterion is recommended when homogeneity of covariances is violated (Hair, Anderson, Tatham and Black, 1998), the significant F value implied that the four band levels for Speaking and Pronunciation significantly differed in terms of the contribution of the individual speech variables to band placement.

4.4 Discriminant analyses

Two discriminant analyses were carried out separately to predict L2 speaker placement in the four Speaking and Pronunciation bands, based on the eight speech measures. These analyses revealed one underlying dimension discriminating between band placements. As shown in Table 10, this dimension (Function 1) accounted for 92.6% of total variance in the IELTS Speaking band placement, canonical $R^2 = .27$, distinguishing across the four Speaking bands, Wilks' $\Lambda = .71$, $\chi^2(24) = 209.159$, $p < .0001$. A single dimension (Function 1) also accounted for 90.2% of variance in IELTS Pronunciation band placement, canonical $R^2 = 25$, differentiating between the four Pronunciation bands, Wilks' $\Lambda = .73$, $\chi^2(24) = 192.271$, $p < .0001$.

Tables 11 and 12 show standardised canonical correlation coefficients representing associations between each speech variable and the significant dimension (i.e., Function 1) in each analysis. These coefficients (collectively referred to as the structure matrix) indicate the relative contribution of individual speech variables to each discriminating dimension (function). Standardising the coefficients ensures that scale differences between the variables are eliminated, and using absolute weights (i.e., ignoring the directionality of the relationship) allows for ranking each predictor variable, such that the variables with large weights are those which contribute most to differentiating the groups. As shown in both tables, all eight speech variables had the largest absolute associations with a single discriminant function (Function 1), and correlation strengths patterned in a similar way in terms of their contribution to separating the IELTS Speaking and Pronunciation bands. Grammatical accuracy and sentence structure, along with lexical richness, intonation and speech chunking, had the strongest association with the discriminating function (Function 1). In contrast, word stress, speech rate, segmental errors and comprehensibility had weaker associations with the discriminating function (Function 1).

Function	Eigenvalue	% of Variance	Cumulative %	Correlation
Speaking band placement				
1	.376	92.6	92.6	.523
2	.024	5.9	98.5	.153
3	.006	1.5	100.0	.077
Pronunciation band placement				
1	.330	90.2	90.2	.498
2	.020	5.5	95.7	.141
3	.016	4.3	100.0	.125

Table 10: Eigenvalues for discriminant functions

Measure	Function		
	1	2	3
Grammatical accuracy & sentence structure	.930	.061	.087
Lexical richness	.861	.247	-.051
Speech chunking	.758	.339	.075
Intonation	.757	.097	.310
Word stress	.674	.047	-.246
Speech rate	.669	.495	-.178
Vowel & consonant errors	.536	.201	.147
Comprehensibility	.474	.408	.078

Table 11: Structure matrix for IELTS Speaking scores

Measure	Function		
	1	2	3
Grammatical accuracy & sentence structure	.945	.061	-.055
Lexical richness	.863	.134	-.014
Speech chunking	.745	-.002	.336
Intonation	.735	.188	.191
Word stress	.668	-.037	-.072
Speech rate	.633	.297	-.348
Vowel & consonant errors	.551	.337	.149
Comprehensibility	.462	.041	.023

Table 12: Structure matrix for IELTS Pronunciation scores

Tables 13 and 14 illustrate the functions at group centroids, which are the mean function scores for each group. Of relevance here is the sign of the centroid (positive or negative), because it indicates which groups have been discriminated from which. Focusing only on Function 1 (the only significant discriminating function), it appears that Function 1 discriminates the lowest bands (Bands 5 and 6) from the higher ones (Bands 7 and 8) for both Speaking and Pronunciation band placements. Notably, Function 1 discriminates more robustly between Bands 5 and 8, because differences between centroids are greatest for these groups.

IELTS Speaking bands	Function		
	1	2	3
Band 5	-.720	-.158	.018
Band 6	-.314	.205	-.080
Band 7	.387	.088	.104
Band 8	.911	-.152	-.078

Table 13: Functions at group centroids for IELTS Speaking scores

IELTS Pronunciation bands	Function		
	1	2	3
Band 5	-.645	-.204	.036
Band 6	-.392	.158	-.070
Band 7	.425	.068	.228
Band 8	.770	-.078	-.119

Table 14: Functions at group centroids for IELTS Pronunciation scores

The same relationship between the discriminating function (Function 1) and band placements is illustrated using a combined-groups plot, shown in Figures 2 and 3 representing IELTS Speaking and Pronunciation scores, respectively. These figures present mean scores for each L2 speaker, grouped according to each IELTS band they were assigned to. Group centroids or mean function scores for each group are designated by squares. Although both graphs plot each speaker in a two-dimensional space (with Function 1 plotted along the *x*-axis and Function 2 along the *y*-axis), it is clear that the Speaking and Pronunciation bands are distinguished by a single dimension (Function 1) along the *x*-axis.
In essence, Function 1 best discriminates between Bands 5 and 6 on the one hand and Bands 7 and 8 on the other, as designated by the black squares on the horizontal plane.

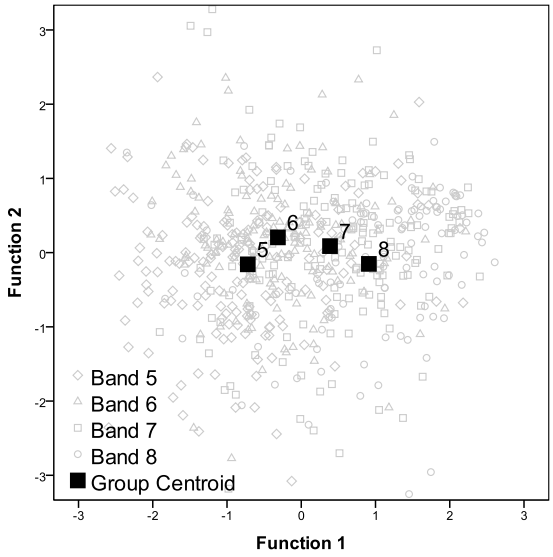


Figure 2: Discriminant function scores for speaking band placements, with mean centroid values designating IELTS Speaking bands 5 through 8

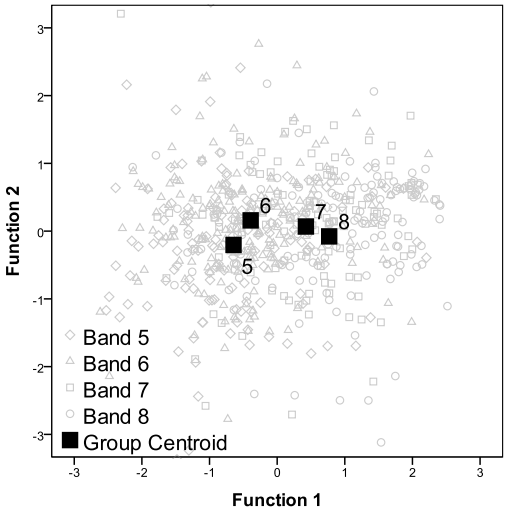


Figure 3: Discriminant function scores for pronunciation band placements, with mean centroid values designating IELTS Pronunciation bands 5 through 8

The final aspect of discriminant analysis involves a cross-tabulation of classification results, based on the significant discriminating function (Function 1). Tables 15 and 16 show the tallies of observed and predicted group memberships (i.e., band placements), separately for IELTS Speaking and Pronunciation scores. When prediction is perfect all cases should lie on the diagonal (designated by grey shading in the tables), with 100% of the total group memberships fully classified into their respective groups through a discriminating function.
The classification results revealed that only 46.5% of IELTS Speaking scores and 47% of IELTS Pronunciation scores were classified correctly. Focusing on the speaking scores, the most accurate classification occurred for Band 5 (72% of the cases classified correctly), while particularly problematic classification was found for Band 6, with 80% of the cases misclassified. For pronunciation scores, the two most accurate classifications were yielded for Bands 6 and 8, with 72% and 63% of the cases classified in accordance with the IELTS band placements, respectively. The worst classification was for Band 7 with a striking 92% of misclassifications. In essence, cross-tabulation results suggested that Band 6 was particularly problematic in scoring speaking, while Band 7 was particularly problematic for scoring pronunciation.

Speaking scores		Predicted group membership				Proportions	
IELTS bands	Total	5	6	7	8	Misclassification	Correct classification
Band 5	174	125	13	31	5	28.2%	71.8%
Band 6	147	71	29	40	7	80.3%	19.7%
Band 7	170	39	23	78	30	54.1%	45.9%
Band 8	116	21	2	43	50	56.9%	43.1%

Table 15: Classification results for IELTS Speaking scores

Pronunciation scores		Predicted group membership				Proportions	
IELTS bands	Total	5	6	7	8	Misclassification	Correct classification
Band 5	137	38	78	0	21	72.3%	27.7%
Band 6	200	24	143	3	30	28.5%	71.5%
Band 7	119	5	45	10	58	91.6%	8.4%
Band 8	151	6	46	4	95	37.1%	62.9%

Table 16: Classification results for IELTS Pronunciation scores

To summarise, discriminant analyses were conducted to determine which of the eight speech measures most effectively discriminated between IELTS Speaking and Pronunciation bands (Band 5, Band 6, Band 7 and Band 8 and over). The semantic differential speech measures used as predictor variables were comprehensibility, segmental errors, word stress, intonation, speech chunking, speech rate, lexical richness and grammatical accuracy. Results overall revealed that group separation in IELTS band placement can best be explained by a single underlying dimension, accounting for 93% of between-group variability in IELTS Speaking scores and 90% of variability in IELTS Pronunciation scores. Closer analysis of individual predictor variables revealed that all speech measures loaded on the same discriminating function, with measures of grammatical accuracy and lexical richness having the strongest associations with the discriminating function and measures of segmental accuracy and comprehensibility having the weakest associations. Finally, cross-validated classifications showed that overall 46.5% of speaking and 47% of pronunciation scores were classified correctly, and that the clearest distinctions between bands were between the combined Bands 5 and 6 and the combined Bands 7 and 8.

4.5 Between-band comparisons for the Speaking and Pronunciation scales

To provide a more comprehensive picture of the relationship between individual speech measures and IELTS speaking and pronunciation band placements, we followed up discriminant analyses with univariate ANOVAs, carried out separately for each speech measure. The goal of these analyses was to determine potential between-group differences in IELTS Speaking and Pronunciation scores as a function of the individual speech measures. The ANOVAs comparing the scores for the eight speech measures across the four IELTS Speaking bands revealed significant *F*-ratios in all cases, as illustrated in Table 17, thus confirming the results of preceding discriminant analyses which suggested that all speech measures contributed to distinguishing between the four groups. Because the assumption of homogeneity of variance (according to Levene's tests) was violated in all cases, between-band comparisons were carried out using Tamhane's post-hoc tests.

Measure	*df*	*F*	*P*
Comprehensibility	3623	19.151	.0001
Vowel & consonant errors	3622	23.047	.0001
Word stress	3623	36.046	.0001
Intonation	3628	23.021	.0001
Speech chunking	3629	46.656	.0001
Speech rate	3629	38.472	.0001
Lexical richness	3626	58.215	.0001
Grammatical accuracy & sentence structure	3627	68.169	.0001

Table 17: Summary of univariate ANOVAs for IELTS Speaking scores

Post-hoc tests showed that the measures of grammatical accuracy and sentence structure, lexical richness and word stress significantly distinguished between all four IELTS Speaking bands ($p < .05$). The remaining measures (i.e., comprehensibility, segmental errors, intonation, chunking, and speech rate) significantly differentiated between Bands 5, 6, and 7 but failed to distinguish Bands 7 and 8. Put simply, the upper speaking bands were discriminated only through measures of grammatical accuracy and sentence structure, lexical richness and word stress. The findings of between-band comparisons are summarised in Table 18, with merged cells representing lack of significant between-band differences.

Measure	Band 5	Band 6	Band 7	Band 8
Comprehensibility				
Vowel & consonant errors				
Word stress				
Intonation				
Speech chunking				
Speech rate				
Lexical richness				
Grammatical accuracy & sentence structure				

Table 18: Summary of between-band comparisons for IELTS Speaking bands

Similar univariate ANOVAs comparing the scores for the eight speech measures across the four IELTS Pronunciation bands revealed significant *F*-ratios in all cases, as illustrated in Table 19, again supporting the results of preceding discriminant analyses. As the assumption of homogeneity of variance (according to Levene's tests) was violated in most cases, between-band comparisons were carried out using more conservative Tamhane's post-hoc tests.

Measure	*df*	*F*	*P*
Comprehensibility	3623	15.649	.0001
Vowel & consonant errors	3622	22.029	.0001
Word stress	3623	30.859	.0001
Intonation	3628	22.593	.0001
Speech chunking	3629	38.937	.0001
Speech rate	3629	30.908	.0001
Lexical richness	3626	51.659	.0001
Grammatical accuracy & sentence structure	3627	62.216	.0001

Table 19: Summary of univariate ANOVAs for IELTS Pronunciation scores

Post-hoc tests showed that none of the eight speech measures distinguished significantly between all IELTS Pronunciation bands. In fact, the clearest differences emerged between the combined Bands 5 and 6 and the combined Bands 7 and 8, with all measures showing a clear distinction between the two lower and the two higher bands. The two lower bands were distinguished only through measures of speech rate and lexical richness ($p < .05$), and the two higher bands were not distinguished by any speech measure. Put differently, between-band comparisons of the eight speech measures yielded little evidence that Pronunciation Bands 5 and 6 as well as Bands 7 and 8 were distinguished through any of the targeted speech measures, thus supporting the results of discriminant analyses. Table 20 summarises the results of the post-hoc comparisons, with merged cells again representing a lack of significant differences between bands.

Measure	Band 5	Band 6	Band 7	Band 8
Comprehensibility				
Vowel & consonant errors				
Word stress				
Intonation				
Speech chunking				
Speech rate				
Lexical richness				
Grammatical accuracy & sentence structure				

Table 20: Summary of between-band comparisons for IELTS Pronunciation bands

5 QUALITATIVE RESULTS

5.1 Comparing the retired 4-point with the revised 9-point Pronunciation scale

At the beginning of the focus group sessions, prior to conducting ratings of speech, examiners were asked about their preference for using the discontinued 4-point versus the revised 9-point IELTS Pronunciation scale, with seven out of eight examiners familiar with the older system. E2, E3 and E6 directly reported a preference for the 9-point IELTS Pronunciation scale due to greater flexibility in assigning scores. E4 elaborated:

> *E4: I feel better using the 9-point in the sense that I feel like I'm giving them a more accurate mark. But the 4-point is much easier because it was just easier to make the decision really. Quite often when I used that one I felt like I needed something in between to really justify the score…Virtually everybody got a '6' under the old scale. They had to be quite bad to get a '4' and very good to get an '8' so it was almost as if you didn't really have to think very much about the pronunciation at all in those days, '6' was a default and if they weren't good enough they were a '4' and if they were better they were an '8'.*

E4's description echoes the finding from Brown (2006) and DeVelle (2008) that Band 6 was often treated as the default level in the old (4-point) Pronunciation scale, which spurred the development of the revised IELTS Pronunciation scale.

E5 offered an alternative perspective.

> *E5: Whenever I rate now I always go back to five years ago when I was marking the 4-point scale and then adjust up or down. I always use that one from four or five years ago. It's always in my mind, that one.*

E5's admission of resorting to his prior experience using the 4-point scale as the initial point of reference for completing his 9-point Pronunciation ratings suggests the possibility that not all IELTS examiners who were trained on the old system have successfully transitioned to thinking in 9-point scale mode for Pronunciation, although he was the only examiner in the dataset who attested to doing this.

5.2 Assessing pronunciation in relation to other aspects of test-taker ability

The lower mean ratings from the examiners' background questionnaire regarding comfort rating Pronunciation relative to the other IELTS Speaking components were corroborated in examiners' qualitative comments, with E3 explicitly stating that Pronunciation is 'the one I struggle with the most…out of the four [Speaking scales]'. Other examiners' comments unveiled throughout the course of the focus group sessions made reference to comparisons with assessing other aspects of test-taker L2 proficiency that were more straightforward, which is the focus of this section.

Half of the raters (E2, E3, E4, E5) attested that Pronunciation is the scale that they rated last in operational testing settings, although none suggested that this was because Pronunciation was ordered after the other Speaking subscales in the IELTS Speaking bands that they consulted (rightmost column).

E3 articulated a counter reason that Pronunciation could be rated first, although never suggested that she did this:

> *E3: [Pronunciation] it's the thing that's least likely to change over the course of the test because the pronunciation isn't going to get better or worse in the 14 minutes that quite often their fluency does as they get less nervous or their grammar gets a bit more sophisticated as they relax into things. It's almost tempting sometimes to think of a mark in the first minute for pronunciation and then start looking for some other things.*

No other examiners described having any such impulse to score Pronunciation first, although E4 independently noted that pronunciation ability was more robust to test-taker's nerves than other rated aspects of speaking proficiency (Fluency and Coherence) and E2 seemed preoccupied with detrimental effects of test-takers' nerves on their speaking performance in general. In terms of arriving at Pronunciation scores after other component scores had already been assigned, the overall approach, particularly expressed by E2 but also echoed by colleagues, was that 'you can rightly or wrongly mark straight by the other scores'.

> *E2: If someone's a '6', you go 6, 6, 6, and then pronunciation and unless it's really good or really bad you kind of give it a '6', unless it's really poor. If someone's a '7' for the other three things, you're thinking '7', '7', '7.' You come to the pronunciation, unless they're particularly good, which is unlikely, or particularly weak because they come from Korea or something where they leave off all the ends of the words, you tend to just think, plump for '7' and maybe that's not the right way of doing it. I think that's what probably most of us do.*

> *E3: I think if their pronunciation is really really poor, then it's easier. So if they're good on the other things and you don't really notice the pronunciation, you do tend to go higher just because the others are higher.*

Examiners described that minor adjustments (within one band) could be made to Pronunciation relative to the other subscales if necessary to yield a jagged profile, although this appeared from their descriptions to be the exception rather than the rule. In terms of how the examiners arrived at the initial band score for any of the Speaking bands, there appeared to be a broad community consensus on what constituted the crucial score of Band 7 from the individual listener's (examiner's) perspective.

> *E4: Well someone who had been doing IELTS examining for a long time once said to me, what I do is I think would I want to go out for dinner with this person and if the answer's yes then they're '7' or above and if it's no they're a '6.' And that's clearly not a sufficient way of doing it but it is the kind of process that's going on in the examiners' minds.*

> *E2: Somebody told me when I started doing this, a person who is '7' is someone you could have dinner with. And you're not thinking, oh this is a foreigner, you're able to talk them, they're making mistakes it doesn't matter but you're at ease, it's comfortable. And I still use that and that comes into the pronunciation because if you kind of go, what's he saying, I don't really understand what he's saying. If pronunciation is a '7' then you relax into it and it doesn't matter if there are grammar mistakes or inappropriate use of vocabulary. So it's ease, a kind of sense of ease and comfort and I mean you couldn't really put that, I'm not saying you could put that in the descriptors but I think that's what we have in our heads.*

> *E7: I think it's something that the trainer told us in the training two years ago. If you could have a conversation then quite happily in the evening, then I'll go for a 7, but that's how I start. It's that realistic kind of. You can't be that mechanical about it, because it's got to be, okay, I'm spending time with this person. If I'm gonna charge £20 by the end of an hour, then I'll go for a 6 or 5. But if I wanna leave the room, then it's kind of, I know it sounds terrible, but that's what it is, isn't it, and then I think, wow hang on, how long have you been here, then it's an 8. It's that impression that that person gives me.*

> *E1: The other thing when you're talking about the general rule of thumb for the bands and you're saying the 7 and going out for dinner, I think the IELTS scale is there really so that people going to university are going to be able to survive and do well on their courses alongside, you know, British, American, native-speaker students. And you're thinking well, if someone is a 7 then or 7.5, they will be able to cope. If they're a 6, they're probably going to struggle. If they're a 5, it's better really for them not to begin to approach, even if they can find a university that will take them on. They're just setting themselves up to fail, and I think that kind of thinking goes through my mind.*

This glimpse into examiners' scoring by gauging their attitudes towards engaging in conversation with the test-taker for a prolonged period of time and extrapolating that to Speaking score assignment appears to have arisen in an IELTS training session, as E7 suggested, and perhaps by the same IELTS trainer in Southern England. E1's comments further underscore examiners' cognizance of the high-stakes consequences associated with assigning test scores around Band 7 for test takers in academic settings. In an unrelated discussion, E8 raised the point that it is not just intuition that governs examiners' judgments. Regardless of an examiners' familiarity with or a particular L1 accent, for example, there is also the crucial component of 'meeting the standard'.

One issue that arose that posed a particular dilemma for some examiners related to the overlap between the Fluency and Coherence and Pronunciation scales.

> *E3: And often you find you mark them down twice because if they're not fluent then they're not using the appropriate chunking so then you mark them down on fluency and pronunciation.*
>
> *E4: I always have trouble in the test distinguishing between coherence and pronunciation. And if there's a problem with chunking I never know whether to choose that in fluency or coherence. And I try not to mark them down twice.*
>
> *E5: Sometimes when they lose their coherence that can affect your rating of the pronunciation scale. So they might have a long extended discourse and they're speaking very quickly but they might be intelligible but it's not coherent so as a rater you're thinking is this fluency or coherence or is this pronunciation?*

'Chunking', what constituted it, or the way it interfaced with different aspects of speech performance (e.g., E8: 'lexical resource and pronunciation…less common idiomatic items or chunks of language that they're comfortable with using…sustainable appropriate rhythm…'), and what was part of the remit of the Pronunciation scale, as opposed to the Fluency and Coherence scale, was also the subject of examiner anecdotes about scoring dilemmas (e.g., test-taker speaks too fluently/rapidly for the examiner to be able to understand every word). It also prompted examiner exchanges, including the following, in which input was sought from focus group colleagues after conducting speech ratings in the research setting.

> *E1: Did you find both of you that speech chunking and speech rate are quite significant elements when it comes to this sort of pronunciation?*
>
> *E2: Speech rate is someone who speaks very fast, but the chunking, I've got a slight problem with chunking, though, because for example let's look at chunking when someone says, 'how are you?' So if you get it wrong you would have to say, 'how... are... you?' But I can't quite see how you would do it any other way. It's when you haven't got brakes if you like.*
>
> *E1: I don't think that's a particularly good example actually if you don't mind me saying so. I mean it's how you just use three words. But when they're trying to put together, you know, quite long stretches of language you notice when they're sort of hesitating in between, in places where, you know, we wouldn't normally hesitate.*
>
> *E2: You're hesitating in the right places so your chunks are right.*
>
> *E4: Yeah and I found that I didn't know what to do about people who were chunking more or less appropriately but not connecting their individual words together so I had, I don't know what nationality he was but he was pausing in all the right places but his words were just too clipped, too individual really, like a German speaker. And there didn't seem to be a category for that. I don't know if you have the same problem or if you just included that in chunking or ...*
>
> *E5: Yeah, I put that as chunking.*

Different examiners reported deferring to different subscales that they had scored previously before deriving their Pronunciation scores.

> *E5: I don't make it [pronunciation] massively different from the rest because fluency and coherence is the one that's more important than the pronunciation.*
>
> *E7: Pronunciation, if I have any doubts, I kind of think back on the lexical resource just to one extent, because I feel it's kind of close together for me.*

A final point that arose in the discussion with regard to assessing other aspects of proficiency was bringing in insights from assessing writing in the IELTS.

> *E5: [Re. comparison of IELTS Speaking subscales] Easiest would be grammatical range and lexical resource [subscales]. The reason why I find those easier is because I also mark writing as well, and they are pretty similar to the writing descriptors. Not the same descriptors exactly, but it's the mix of short and complex sentences. You take a risk but you don't succeed, you take a risk with the vocabulary but you don't succeed, and that gives you the benchmark for a '6'. And that gives you a benchmark for the speaking and the writing, so they're quite easy for me to go straight into them. I know what they are.*
>
> *E4: Is there a reason why the writing descriptors have ceilings and the speaking ones don't? Because writing has things like, if you don't use paragraphs you can't get more than '5' for example. And that makes it very easy for examiners to rate those but there's nothing like that in the speaking.*

E5's initial musing about the facilitative effect of having parallel bands for Writing and Speaking was followed up by E4's comment on the possibility of having capped features for assessing speaking. This notion was imported into a subsequent focus group session, held on a different day with a new mix of examiners.

> *E3: There are ceilings in the writing descriptors and there's none in speaking. Yeah it does make it a lot easier. I was wondering if you had something like you know, for level 7, I don't know how you would word it, but you know, uses the features of connected speech…If you knew that if you had to do that to get a 7, I think it would give more consistency amongst the examiners because it would give you something much more specific to sort of listen for without actually putting too much of a load on examiners.*

E4: So in the writing, for example, if they haven't used paragraphs, then they cannot get…

E3: Yeah they can't get more than a 5 for coherence. And is it framed positively in the scale or...
Well it's not, some of them are positive and some of them are negative, aren't they. So for task achievement in the writing, to get a 6, you have to have a clear overview.

E4: But I'm sure there's more consistency because we know that you cannot give a Band 6, unless there's an overview in task 1.

E3: It just gives a much clearer idea to the examiners of what IELTS considers are the most important features.

E1: Well, I mean those kinds of criteria objectively are sort of assessable, aren't they, I mean task achievement. If someone included that initial sentence with an overview, it's there or it's not there. But with speaking, it's more subjective I think.

The examiners' words arguably speak for themselves in describing the facilitative effects that identifying 'ceilings' would have on their scoring. However, E1's acknowledgment that this may be more difficult to implement for the spoken rather than the written medium is likely, in part, due to the ephemeral and intangible nature of speech (Isaacs, accepted). Naturally, the IELTS examiners would need to establish the presence or absence of the stated feature in real-time during the course of the Speaking exam if this recommendation was to be followed. Leaving this consideration aside until the Discussion, the next section highlights examiners' perspectives on interpreting key terminology in the IELTS Pronunciation band descriptors.

5.3 Terminology used in the IELTS Pronunciation scale

This section will focus on IELTS examiners' views on rating 'phonological features,' 'comprehensibility' (termed 'intelligibility' in the IELTS Speaking band), and, finally, the descriptor, 'shows all the positive features of Band X and some, but not all, of the positive features of Band Y', particularly in reference to IELTS Pronunciation Band 7, which is often important for consequential decision making in higher education settings.

5.3.1 Phonological features and nativeness

Early on in the focus group session when conversing about the scale descriptors prior to rating speech samples, E2 pondered, 'the adjectives you know, what's a full range compared to a wide range?' in reference to the qualifiers for 'phonological features' at Bands 9 and 8, respectively. A few conversational turns later, with her query remaining unanswered by her focus group colleagues, she addressed her own question.

E2: Well I suppose in that particular question, a full range is, I always think of a native speaker. But the problem with that is of course that native speakers have different ranges but the average native speaker has a full range, and so a wide range if they're not quite as fully developed.

Thus, 'full range', in E2's interpretation, serves as a means of indirectly evoking the notion of having an educated native speaker at the top end of the scale, although she did acknowledge a degree of variability in native speakers' development or productions. This is echoed in another of her quotes later on in the session.

E2: If you go back to the '9'. Let's assume that all of us here [examiners] are a '9' because we're native speakers and we're all hesitating and we're all different things and some of us speak faster than others. We're doing this, whereas an '8' has some occasional lapses and to me they sound foreign if you like, you know, it's not quite English. So they can do pretty well everything despite occasional lapses and then it kind of goes down, doesn't it until you get right down to the bottom where they can't really do any of this. Well I couldn't if I was doing it in Chinese.

However, E2's notion of the native speaker being at the top end of the scale contrasts somewhat with the views of her colleagues.

E7: I don't think accents really should play any part until they start interfering.

E1: I think most people operating in our field are kind of comfortable with the notion of World English and English being used across lots of areas of the world, and we're tolerant of the idea of there being different accents. So the accent itself I don't think clouds our judgment until it affects intelligibility. And then the minute it affects intelligibility then we're beginning to think, hahah, another native speaker. Not only me the examiner might have problems understanding what this student is saying. Therefore, it goes down a notch in the pronunciation scale.

E2's view is a clear articulation of Levis' nativeness principle (2005)—the notion of native-like perfection, in this case, being the target against which learners should be judged at Band 9 and which distinguishes them from a speaker who apparently 'sound(s) foreign' at Band 8. This view contrasts with the views of E7 and E1, who appear to espouse Levis' intelligibility principle, or the notion that accents only pose a problem when they impede listener understanding. By implication, in a rating context, not all perceived deviance from a native speaker norm needs to be penalised (i.e., only when it interferes with comprehensibility). Thus, even though the native speaker is not directly evoked in the IELTS Pronunciation scale, it is possible to extrapolate that the existing descriptors and under-specification of 'a wide range of phonological features' could be used to support or justify an individual examiner's nativist interpretation or otherwise.

Apparently unsatisfied with her own explanation of nativeness and still probing, E2 persisted to verbalise what was meant by 'wide range' as the focus group session continued. This time, she was successful at eliciting responses from her colleagues and spurred a multi-turn exchange, an excerpt of which is shown here.

> *E2: Well, shall we just look at what is a wide range of phonological features? For example what exactly does this mean? Can you give us an example of someone who's using these features? I mean rhythm is clear, you know, to keep talking in a nice rhythm and then you put stress and intonation. But how are we gonna look at this wide range of phonological features to convey precise and or subtle meaning in pronunciation? Because that's often done in the others things, the lexical resource, isn't it?*
>
> *E3: Yeah, I don't think that first sentence is very helpful at all really.*
>
> *E1: If you were to break it down too much, though, and be kind of very very descriptive in terms of what that actually means, you'd end up complicating the life of the examiner, I think, because it's quite difficult to appreciate that we're kind of administering the exam and assessing the students against already four sets of criteria. I think if you were to expect us to sort of tick boxes within that last pronunciation column along the lines of, I don't know, ability to assimilate, ability to link words correctly in connected speech and so on, we'd be hard pushed.*
>
> *E2: Yes, I agree with you, I don't think I really… consider that sentence much. Not sure I totally understand it even.*
>
> *E3: Well in band 8 it's kind of redundant isn't it. Because the next paragraph talks about rhythm and stress and intonation and accent and intelligibility, so what else?*

E2's suggestion that pronunciation may not be nuanced enough to convey precise or subtle meaning in the first part of that exchange was challenged by E1 a few turns after the excerpt shown above, who claimed that, 'holding the attention of the listener' through pausing and intonation could represent that level of nuance. E1 also made the point in the passage that examiners would be overloaded if they were to have to assess more atomistic pronunciation-related criteria in real-time during a live test. Finally, in the last turn, E3 arrived at the conclusion that the 'full range of phonological features' descriptor was redundant. On this point, she subsequently explained:

> *E3: I was looking at Band 8 particularly because it actually, in the second paragraph, it talks about rhythm and stress and intonation and then it talks about accent and intelligibility. And I mean to me those are quite a lot of phonological features. What else are we talking about?*

In other words, in cases when the list of possible phonological features that could come into play is provided in the descriptors, reference to phonological features could perhaps be omitted. E6 later suggested that 'phonological features' could be interpreted as a heading (e.g., for features such as chunking, rhythm and stress-timing). However, she pointed out a potential caveat.

> *E6: Well it sometimes seems to me that the feature descriptions are, they're like a cluster of characteristics, they don't always go together. I mean you have one person who can use a range of phonological features…But individual words and phonemes may be mispronounced, but it only causes occasional lack of clarity when maybe those two don't go together. You can use the suprasegmentals well and the segmentals not well.*

From the above exchanges, it appears that examiners were confused by the generic term 'phonological features', the qualifiers and further specification of what those features might include within the descriptors. Some examiners viewed it as not contributing valuable content and suggested its omission.

The next section reports on examiners' impressions of the descriptors for Pronunciation Bands 5 and 7, which describe the pronunciation features only as being somewhere in between the adjacent Pronunciation band descriptors in terms of observed features in the test-taker's performance.

5.3.2 The in-between IELTS Pronunciation band descriptors

One point that was robust in the dataset was examiners' observations of the greater time and processing demands involved in Bands 5 and 7 not having unique (independent) descriptors. This is because they needed to consult adjacent band levels to arrive at a determination about a test-taker's IELTS Pronunciation level.

> *E6: When you're constrained of time, you actually have more to read. And I think that's sort of a very basic thing, isn't it, because you look at 6 and then you look at 8, and it takes you a lot more time.*
>
> *E8: I mean if you were literally doing that every time it would take longer.*
>
> *E3: I think it's quite time consuming to actually do when you look at that because then you have to go back and read all the positive and work out which ones are the positive. I mean colour coding would be helpful because you can just see which are the positive features at a glance. I mean obviously you have to spend time reading the descriptors anyway, but I do find each time I'm thinking about a '7,' it is quite time consuming.*
>
> *E2: This is quite confusing when you get here, and you've got to stop and you're going, the positive features of '6' but not, that is actually kind of hard. It's time consuming and it's difficult because you gotta read both [descriptors].*

E7: It's really hard just to read it all, and then which ones do you take? You know I've had to listen for fluency, the lexical resource or grammar, so it's really hard to kind of remember everything. So for me the pronunciation, the effect it has on the listener, whether we did have to strain or there was a rhythm that kind of, not disturbed but I notice. So that's why I tend to look at the negative aspects and see how much they affect me. I've kind of worked the other way round, rather than looking at the positive, I look at the negative because then I'll notice those more.

Clearly, examiners viewed consulting two band descriptors and reflecting on whether pronunciation features from the performance sample were present or absent was time and attention demanding in real-time during the test. Examiners also commented on the imprecision of the wording and on the scope accorded to the examiner in how rigidly they applied the criteria.

E3: But obviously there's quite a huge difference between, 'can be easily understood throughout' and 'can generally be understood throughout'. I think if they put a sentence in 7, then it would become a bit woolly…I think it's nice that there is that leap because then you can just work out what goes in the middle. But I think if they use too many words to describe what's in the middle, do you understand what I mean?

E1: Yeah, it would be difficult to find a word that would sort of become in between the two.

E3: Yeah, it would be difficult to find a word, whereas it's quite easy to imagine what's between easily and generally. And the fact that there's nothing in the middle is fairly easy for me to fill it in.

E8: It's sort of wordy, it's sort of imprecise. I do find it difficult, that you're…trying to sort out what actually is in 7, in the box. What should be there, what would the wording be in there? So I find it very unhelpful.

E1: I think the points in the scale where it says, 'displays all of one but not all of the other' is a bit of a cop-out actually. It's not a clear descriptor. But having said that, I think we probably get a sense of people probably being on a '7' as opposed to an '8' or a '6,' so possibly it's ok.

E6: There's an awful lot of imprecision there. So when you then say displays all of the positive features, does that mean all, generally, or some?

E4: It kind of leaves a lot of space for impression, you know, of the examiner, because some people might be stricter on this [displays all positive features] than others.

E2: You've got to have all the positive features so it's got to be at least a '6' then the question is, is this person an '8'? And if they're not quite an '8' then you put a '7' so that is a bit of a wide, do you know what I mean? There's quite a jump there.

E3: Actually 6 is a large band, isn't it. And perhaps they're just above a 5 or they're almost at a 7.

Although examiners' remarks about the vagueness of the in-between band descriptors were not uniformly negative (e.g., E3), the ambiguity and latitude accorded to examiners in interpreting whether all of the criteria described in the lower band had been satisfied was not in dispute. The final section of the Qualitative Results turns to the way in which examiners interpreted the comprehensibility criterion (termed 'intelligibility' in the terminology of the IELTS Pronunciation scale).

5.3.3 Comprehensibility

One consideration cited in the examiners' comments that underpinned an examination of assessing speaking in general and comprehensibility in particular in the current study was the variable sound quality of the Cambridge English speech data that they rated for research purposes. Examiners framed their impressions in the following ways:

E2: I missed bits then I thought is it the speaker's fault? Is it the recording's fault? Is it my fault? Yeah we do try to be fair.

E4: There's quite a variety of recording quality. Some of them sound like they're inside the womb. Some of them were perfectly clear and others you really had to concentrate to hear, and the ones that's really clear, you just relax. Well then I gave ratings to all of them but sometimes I felt that I was being unfair, really, because there's no way that I can give a proper rating.

E7: I found I had problems with the quality of the recording, especially for the vowel-consonant sounds and I think at times I wasn't sure if I couldn't understand because of the quality or it kind of made it sound a bit flatter, muffling, yeah and so I think it did definitely affect my understanding.

This limitation in their reported ability to hear individual sound files notwithstanding, examiners were still able to comment on dilemmas in scoring comprehensibility. As has been highlighted in Isaacs (2008) and Isaacs and Trofimovich (2012), intelligibility/comprehensibility have been defined and operationalised in a multitude of ways in the L2 pronunciation literature and in assessment instruments. Divergent interpretations of the construct are shown in the italicised quotes below, and the authors' interpretation is provided in the non-bolded unitalicised content in parentheses immediately following each quote.

E8: You can work out words (emphasis on understanding every single word).

E6: Difficult to understand their meaning, not to understand exactly what they're saying (tension between meaning- vs. word-based notions of comprehensibility).

E2: It's the effort. So a '4' is 'understanding requires some effort' and it actually hurts almost when you're almost trying to listen, whereas a '6' 'can generally be understood throughout without much effort', so you can listen, and the mistakes they make don't cause you to have that strain (closely corresponds to the notion of 'painstakingly effortful to understand' used in the semantic differential scale for comprehensibility in the current study; Appendix 4).

E7: The ease of which, how easy it is to understand that person, how much effort they're making, because sometimes they'll catch themselves, you know, what was that word (definition oscillates between the listeners' effort in understanding and the test-taker's effort in producing accurate utterances and self-correcting).

E1: Some of the um Arabic speakers I listened to today, you know their 'r's are sort of still ringing in my ears. You know but it's not an error that really prevents comprehension, so very much a feature of their speech there. So I think it's kind of as I was saying earlier on, it's this ability to make yourself understood that's the key, the key factor, in pronunciation (acknowledgment that perceptually salient segmental errors do not always impede understanding and that it is getting the message across that is most important).

E7: I thought for me there's a difference between effort from the listener and effect on listener so sometimes I could understand them but maybe the intonation or word stress after a few minutes kind of grate on the listener, I don't know what the specific word is, some kind of intonation patterns or word stress, after a while you can start losing the meaning (irritability can be distracting for the listener and can result in difficulty understanding meaning in extended speech).

E8: Would it be comprehensible to a sympathetic native speaker? You stop someone on the street and ask them the way, and if it's a friendly person who's got time, they will offer their ear, although that may not necessarily be the case. And obviously we're tuned in to pick out all the positive things we can, and negative as well. But even if we give a low grade for comprehensibility in the end, we'd still be giving some credit for the other language features, wouldn't we, whereas in a real situation, we're trying to assess their language competence that might not be the case. (acknowledged importance of comprehensibility in real-world communication and of listener factors in making judgments).

E6: Yeah but if we don't look at these discrete features then it's more likely that what we're used to is going to influence your... (need to consider more discrete features to neutralise listener accent familiarity effects that could colour examiners' perceptions of comprehensibility).

E8: I've had examples of people with certain intonation patterns, and I could actually understand it but I know that is really difficult to understand. So it's not so much me, it's my impression, but they're honed by this, by the criteria here. I have understood this person because I've lived in that place and… we've all got different experience haven't we, as examiners, but we all have to, it's the standard isn't it, looking at that, meeting standard (examiners' understanding of speech for learners from a particular L1 background is inevitably affected by their experience, including L1 exposure effects; however, they still need to meet IELTS standards).

Several examiners emphasised the importance of comprehensibility as a key criterion in the scale that governed their decision-making.

E3: I like what it says, the last sentence where it says, 'often unintelligible, understanding required, some effort, can generally be understood throughout without much effort' and those are easy to work with. And actually a lot of the time, yes it is key. I think that's what influences you first.

E5: Well, I think out of those [intelligibility] is the most important pronunciation descriptor. If we're looking at '9', 'can be effortlessly understood,'8' would be 'can be easily understood' and '6' 'can generally be understood'. Those are really benchmarks. Again, you're looking at a holistic feel to their production of speech. I mean, you can get into the technical aspects, but it's a very subjective thing. Those are very important, 'effortlessly', 'easily' and 'generally'. That's quite sensible when you listen to somebody. It should be at the top. I'm surprised, actually, I'm looking at it now I didn't realise it was at the bottom. It should go from general to specific, shouldn't it?

E5 concludes by arguing for a different ordering of the criteria in the Pronunciation scale band descriptors, with comprehensibility listed as the first feature. This point was independently echoed in subsequent focus group sessions.

E7: For the other bands, criteria, I go to the top line, but with pronunciation I realise I'm going to the bottom line, which is determining the grades. So that's the most important, so how much effort does that require me. So that becomes the bottom part, becomes the most important.

Thus, one recommendation is that the sequencing of the criteria within the IELTS Pronunciation band descriptors should be amended, in that ease of understanding (regardless of whether termed intelligibility, comprehensibility, or effortful/effortless understanding) should be described first in the scale bands followed by the more discrete linguistic criteria. As per E5's quote above, this would more closely coincide with examiners' processing of going from holistic to atomistic when listening and making their judgments. This ordering would also coincide with the sequencing of tasks in research studies, including the current study, which involved obtaining listener-scored measures of comprehensibility first followed by more fine-grained measures (see also Isaacs, 2008; Saito et al., 2015).

It was also noted that in the 'Fluency and coherence' scale, test-taker effort is listed as the first element in Band 7 ('Speaks at length without noticeable effort or loss of coherence', public version of the scale, p. 18, IELTS, 2012). It would be logical to follow this order in the IELTS Pronunciation scale for listener effort followed by more atomistic features, either with or without the inclusion of 'phonological features'.

As with the final questionnaire responses on the most important linguistic influences on their understanding of the test-takers' speech, examiners' qualitative comments revealed different personal orientations toward attending to pronunciation-related features in relation to comprehensibility. That is, their comments were mostly idiosyncratic, with no strong group-level finding emerging in terms of the pronunciation features perceived to be the most strongly associated with Pronunciation band level distinctions or that were invariably identified as influencing examiners' understanding. Because the factors that were most important for comprehensibility arose incidentally in the focus group discussions and were not systematically probed in open-response items in the research instruments, they are beyond the scope of the qualitative results reported here. Interested readers could consult research that more directly links raters' perceptions to measures of intelligibility or comprehensibility (Trofimovich and Isaacs, 2012; Zielinski, 2008), which remains a major subject of investigation in current L2 pronunciation research (Isaacs, 2014).

The quantitative and qualitative results of the study are brought together in the Discussion section, which summarises the major findings, discusses research contributions and limitations, and reflects on ways of improving the clarity of the IELTS Pronunciation descriptors for accredited IELTS examiners.

6 DISCUSSION

6.1 Summary and discussion of the main findings

This mixed-methods study examined the linguistic criteria that most efficiently distinguish between upper levels of the revised IELTS Pronunciation scale (bands 5 to ≥8) and how accredited IELTS examiners perceive and engage with the descriptors. As in previous studies, examiners reported less comfort rating the pronunciation scale relative to the other component scales for IELTS Speaking (Galaczi et al., 2012; Yates et al., 2011), lending weight to the need to examine the functioning and examiners' impressions of the Pronunciation descriptors in greater depth in the current study.

The methodology of using semantic differential scales to assess both comprehensibility and discrete linguistic features was piloted on unrelated L2 speech data from an earlier study that had previously been analysed using researcher-coded auditory and instrumental speech measures (Isaacs and Trofimovich, 2012). Following moderate to strong correlations between the rated semantic differential measures and the more objective researcher-coded measures, the semantic differential scales were adopted in the current study. The impetus for this methodological innovation was that the audio files of IELTS test-takers' performances that Cambridge English had provided along with pre-rated IELTS scores were of too variable and, in some cases too poor audio quality to orthographically transcribe and use to elicit researcher-coded measures. However, IELTS examiners were able to provide ratings despite the non-optimal recording quality for the set of speech files, as was evidenced by the existence of the pre-rated speech data.

Eight accredited IELTS examiners in England provided ratings of 80 IELTS test-takers from different L1s performing the IELTS long-turn task (task 2) using the IELTS Speaking band descriptors. They then provided ratings for eight linguistic features on separate semantic differential scales (comprehensibility, segmental errors, word stress, intonation, speech chunking, speech rate, lexical richness and grammatical accuracy-sentence structure). Discriminant analyses using the pre-rated IELTS Speaking and Pronunciation band placements as grouping variables revealed that a single underlying dimension explains between-group variability for both IELTS Speaking and Pronunciation scores (≥90%). Grammatical accuracy and lexical richness were most strongly related to the discriminating function and segmental accuracy, comprehensibility, and intonation had the weakest associations with both scales.

Cross-validated classifications showed that 46.5% of IELTS Speaking and 47% of Pronunciation scores were classified correctly, and that the clearest distinctions arose between the combined Bands 5 and 6 and the combined Bands 7 and 8 and above.

The trend for Speaking, when looking at the individual bands, was that Band 5 was correctly classified the majority of the time (71.8%) followed by Band 7 (45.9%), Band 8 and up (43.1%), and then Band 6, which was very low (19.7%). Although the IELTS Pronunciation scores were equally weighted with the other three Speaking subscales and, hence, accounted for one quarter of the overall IELTS Speaking scores, the classification trend for the IELTS Pronunciation score was very different. Correct classifications were highest for Band 6 (71.5%) and Band 8 and above (62.9%), which had elaborated Pronunciation descriptors. Conversely, classification scores were lowest for Band 5 (27.7%) and particularly Band 7 (8.4%), which happen to be the bands that feature the Pronunciation descriptors, 'shows all the positive features of <the scale band immediately below> and some, but not all, of the positive features of <the scale band immediately above>'.

In their focus group comments, the IELTS examiners underscored the ambiguity of these in-between Pronunciation band descriptors, the interpretative latitude provided to examiners, and the time-consuming nature of locating and consulting the positive features described in the two adjacent bands and relating those to the performance sample in order to make a scoring decision.

One caveat of performing the discriminant analyses in this study from a statistical perspective is that even though the groups were mutually exclusive and collectively exhaustive (i.e., all cases were placed into a group based on the score that had been assigned), the differentiation between groups was not natural in the sense that the IELTS bands are essentially continuous, not categorical (nominal) variables. This notwithstanding, classification accuracy was lowest for the IELTS Pronunciation between-band levels 7 and 5. Further, examiners expressed considerable difficulty applying these descriptors given attentional constraints and under the time pressure of operational examining situations.

A practical recommendation that follows is that the Pronunciation descriptors at Bands 5 and 7 should delineate specific pronunciation criteria in order to implement a clearer division between the groups and to lessen examiners' cognitive load of needing to consult multiple descriptors to arrive at a scoring decision. However, some examiners were reticent about the idea of introducing qualifiers between, for example, 'can be easily understood' and 'can be generally understood', which they cautioned would not elucidate the degree of understanding at the intervening level. Others advised that a checklist of discrete features would not be manageable for examiners during real-time testing. In addition, issues of generalisability of the criteria to learners from diverse L1 backgrounds could arise if the wording was too specific (Crowther et al., 2014; Isaacs and Trofimovich, 2012). These points highlight that any revisions to scale descriptors need to find that elusive happy medium between being too specific and too generic and also to take into account considerations of the end-user's cognitive processing when applying the instrument.

The univariate ANOVAs revealed that no single individual feature, as measured using the semantic differential scales, significantly distinguished between upper bands of the IELTS Pronunciation scale. In addition, IELTS examiners were not uniform in the linguistic criteria that they identified as being most attuned to in their listening and that was most consequential for comprehensibility. Although it would have been desirable from a research perspective to have identified linguistic features that were uniquely responsible for discriminating between upper levels of pronunciation, this may have been an unrealistic expectation. Pronunciation performance as measured in the IELTS test is complex. With performance samples for learners across numerous L1 backgrounds represented in the study, it is perhaps unsurprising that no single linguistic variable was able to effectively discriminate between the different pronunciation levels.

Pronunciation, and comprehensibility in particular, is subject to L1 specific effects (Crowther et al., 2014; Derwing, Thomson and Munro 2006), and it may be difficult to observe incremental differences across a relatively narrow proficiency range when single measures are being used as predictors. Identifying clear-cut discriminating criteria was an easier task in the Isaacs and Trofimovich (2012) study because the sample consisted of only one L1 (French), the L2 ability range was wider, the linguistic measures were more numerous and more varied, and fewer levels needed to be differentiated (three) than in the current study (four). There is a need for systematic research to determine which pronunciation criteria in rating scale descriptors are universal and cut across L1 background, and which are L1-specific (Isaacs, submitted). It may be that a suite of pronunciation variables, in conjunction with other linguistic factors (e.g., rhythm and lexical choice), work together to feed into band level distinctions. The way that the linguistic variables cluster together could inform future revisions to the IELTS Speaking band descriptors and the Pronunciation scale in particular.

One way that the revised Pronunciation scale could be improved, based on insights from the current study, is to more clearly define the terminology used in the scale descriptors. The glossary in Appendix 5 was developed by the researchers to help the EAP teachers in the preliminary study and the IELTS examiners in the main study interpret the terms used in the semantic differential scales, which, in turn, were partially based on terminology or concepts from the IELTS Pronunciation scale. These definitions were devised in the absence of publically available definitions for the IELTS rating scale criteria and, thus, may not align with the definitional interpretations that the IELTS test developers had intended. Qualitative data that emerged incidentally in the focus group discussions revealed some confusion around terminology in the IELTS Pronunciation scale, with some discussion centring on how performance features that were present in the speech samples related to terms such as 'chunking'.

'Phonological features' was emphasised by several examiners as another term that could benefit from greater definitional clarification. One examiner understood the qualifier associated with this term at Band 9 ('full range') to imply the presence of accent-free, native-like pronunciation, although no other examiners framed their understanding in this way. It would be useful to clarify what the expectation for manifested phonological features is at each level of the scale. For example, can a foreign accent be detected at the top level of the scale, or does the presence of a perceptible accent preclude performance at the highest level of the Pronunciation scale? Related to this, several IELTS examiners recommended following the example of the IELTS Writing scale and identifying pronunciation performance features that need to be minimally present at each band level to achieve the corresponding score. However, it is unclear how feasible it might be to do this in the spoken as opposed to the written medium. Ongoing work on the English Profile (Hawkins and Filipovic, 2012) could perhaps expand the focus to include pronunciation to explore this possibility.

Examiners' comments also exposed numerous interpretations of 'comprehensibility' (termed 'intelligibility' in the scale). Clearly specifying for examiners whether comprehensibility relates to listeners' understanding of every word that the test-taker utters, to their understanding of the overall message, or to the processing effort entailed in sustaining attention to meaning would be beneficial for construct validity reasons (Isaacs, 2008; Isaacs and Thomson, 2013). Examiner familiarity effects, a rater characteristic that has the potential to bias assessments of L2 speech (Winke and Gass, 2013; Winke, Gass and Myford, 2013), also arose in the focus group discussions in relation to comprehensibility. Some IELTS examiners suggested that one way to mitigate between-examiner variability in terms of their exposure to different L2 accents was to focus on test-takers' performance in relation to the specific pronunciation features described in the scale, as opposed to their overall impressions of comprehensibility (however defined). Some examiners additionally underscored a lack of guidance on whether to judge comprehensibility from their own personal perspective as an experienced teacher and examiner, from the perspective of a patient or impatient lay listener, or from the perspective of whether or not the test-taker would likely succeed at university from an oral communication standpoint (i.e., in view of the gatekeeping mechanism that the IELTS normally serves for getting into university). Such issues could perhaps be explicitly discussed in rater training and more carefully formalised in written material available to IELTS examiners and the general public (e.g., language teachers, researchers, test-takers) to enhance their understanding of what appear to be fundamental considerations to the construct of Pronunciation in the IELTS (IELTS, 2007).

Another suggestion for improving the IELTS Pronunciation scale descriptors that arose in the focus group discussions related to reordering the descriptors within each band so that comprehensibility appears first and the more discrete features are listed thereafter. Several examiners reported that comprehensibility was either the superordinate criterion for their IELTS Pronunciation decision-making, or the first element they generally attended to when scoring the speech. A few examiners also voiced that 'phonological features', in its current incarnation, does not add much content at certain levels of the scale, in light of the list of more detailed features likely to be observed at that level, and could be omitted. These possibilities could be explored in further Pronunciation scale validation research.

One final recommendation relates to the poor quality of the audio recordings and the link with operational exam conditions. In the case of recordings that had considerable background noise, it was apparent that some testing environments at international test centres are quieter than others. One IELTS examiner (E5), who now works as an IELTS trainer and has served as an examiner in numerous international settings, made this point in an informal (unrecorded) conversation with the researcher about a year after data collection had been completed. In an ideal world, all IELTS Speaking tests would be completed in a sound-attenuated room, as background noise can be distracting to both the test-taker speaking and responding to listening prompts, and to the examiner conducting the assessment. If comprehensibility is used as a criterion in the scale, live test performances, to the extent possible, should be carried out in environments where the background noise is likely to be minimised. Background noise has been shown empirically to degrade listeners' understanding of L2 speech (Munro, 1998), including in speech that is otherwise perfectly understandable. Therefore, eliminating noisy test conditions in operational testing settings would be desirable in the interests of fairness if comprehensibility and pronunciation accuracy are among the assessed criteria.

The final section of this report addresses methodological considerations in the current study that constitute acknowledged limitations related to the rating procedure. These should be taken into account when interpreting the findings and for the benefit of future research.

6.2 Limitations related to the rating instruments and procedure

The innovation of using semantic differential scales in the current study arose from the necessity of eliciting measures of discrete linguistic features to examine their efficacy in discriminating between upper levels of the IELTS Pronunciation scale but not being able to use the objective measures from Isaacs and Trofimovich (2012) due to the variable sound quality in the test-takers' audio recorded IELTS performance samples. The EAP teachers in the preliminary study applied the semantic differential scales reasonably consistently (intraclass correlations: >.9), which is similar to the result obtained in subsequent research, which made use of slightly modified (non-IELTS influenced) semantic-differential scales via a computer application (Crowther, Trofimovich, Isaacs and Saito, 2015; Crowther et al., 2014; Saito et al., in press).

These scales were sensitive enough to capture L1 effects and task effects in these studies. However, contrary to expectation, the IELTS examiners in the current study did not use the semantic differential scales as consistently as the rater groups in these studies, who were experienced teachers but not accredited IELTS examiners. Indeed, the IELTS examiners' intraclass correlations ranged from .54 for intonation to .80 for lexical richness, which was much lower than the Canadian EAP teachers' internal consistency using identical scales.

Through focus group discussions, it emerged that some IELTS examiners felt constrained by the 5 cm lines used for the semantic differential scales in that they were not long enough to allow them to distinguish precisely enough between speakers who were at a similar ability level. A 10 cm line (or computerised adaptation) of the semantic differential scales would have enabled more space in which to record their scores but was not incorporated in the current study due to space efficiency reasons in the instrument used to record their ratings (Appendix 6). A specific behaviour that was unexpected that likely contributed to the inconsistency was that, whereas the Canadian EAP teachers treated the semantic differential scale as percent scales, several IELTS examiners reported trying to represent the nine IELTS band levels on the 5 cm semantic differential line, although they were in no way advised to do so in the instructions and it was not clear how self-consistent they were in doing this. This behaviour is likely a by-product not only of the sequencing of the rating tasks in the research procedure for IELTS examiners (i.e., rating using the IELTS Speaking band descriptors followed by rating using the semantic differential scales), but also suggests the influence of their extensive experience using and being socialised into the IELTS rating system on assigning scores using other assessment instruments (Barkaoui, 2010; Lumley, 2005). Piloting the procedure on IELTS examiners, who are clearly different than non-IELTS trained teacher raters, may have brought these problems to light. Unfortunately, piloting on IELTS examiners was not pursued in the current study due to the desire to include all raters who had volunteered (limited volunteer pool) in the main study.

Another problem was that the researchers interpreted 'speech chunking' to incorporate the notion of appropriate speech rate, pausing at logical junctures and rapid (automitised) access to pre-fabricated chunks during real-time communication (Segalowitz, 2010). In light of perceived overlap of this construct with rhythm and stress timing and in order to keep the number of semantic differential scales to a manageable number (more than eight would have been difficult), 'rhythm' was not included as a separate semantic differential measure. In retrospect, this decision was an oversight. Several IELTS examiners noted this omission during the focus group debrief at the end of the study. Some incorporated the notion of rhythm (typically measured at the phrasal or sentential level) with word stress (typically measured at the word level), whereas others considered rhythm to be part of intonation.

Other omissions that some IELTS examiner noted in the semantic differential scales when summarising influences on their ratings at the end of the session included linking and the use of cohesive devices (the latter of which seemingly falls under the IELTS Fluency and Coherence subscale).

Although intention was to capture the factors found to be linked to comprehensibility from the Isaacs and Trofimovich study (2012), the semantic differential scales were not comprehensive enough in capturing possible pronunciation-specific influences, which could have been prioritised over the lexical and grammar focused semantic differential scales. Finally, scalar endpoints for the grammar semantic differential scale were, 'grammatical accuracy is poor and/or sentence structures are simple or fragmented' and 'grammatical accuracy is excellent and/or sentence structures are suitably complex'. However, merging these two aspects of grammatical accuracy and syntactic complexity within a single scale represents a possible confound. As one IELTS examiner attested, the simpler the syntactic structures that are used (i.e., less risk on the part of the test-taker), the more accurate the L2 speech might be. These concepts could have been more effectively measured in separate semantic differential scales, although, again, the proliferation of scales was not feasible with the allotted timeframe for the study.

A final methodological limitation is that, whereas the Cambridge English pre-rated IELTS speech samples were scored based on examiners' impressions of performance on all three IELTS Speaking tasks, the eight IELTS examiners in the current study based their ratings solely on the IELTS long turn-task. As one examiner suggested in the focus group discussions, this task tends to be less discriminating than the more interactive task with the interviewer (task 3).

Future research could examine ways of operationalising the constructs in the semantic differential scales in a way that is amenable to measuring both interactional performance, and performance on monologic tasks, such as the long-turn task that was rated in the present study. Due to the small number of studies, to date, on the revised IELTS Pronunciation scale, the potential for future research that builds on the current study and avoids the methodological issues described here is vast.

7 REFERENCES

Alderson, JC, 1991, 'Bands and scores' in *Language Testing in the 1990s: The Communicative Legacy*, eds JC Alderson and B North, Macmillan, London, pp 71-86

Bachman, LF, 2000, 'Modern language testing at the turn of the century: Assuring that what we count counts', *Language Testing,* vol 17, pp 1-42

Bachman, LF, 1990, *Fundamental considerations in language testing*, Oxford University Press, Oxford

Bachman, LF and Palmer, AS, 1996, *Language testing in practice*, Oxford University Press, Oxford

Barkaoui, K, 2010, 'Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study', *TESOL Quarterly,* vol 44, pp 31-57

Berry, V, O'Sullivan, B and Rugea, S, 2013, *Identifying the appropriate IELTS score levels for IMG applicants to the GMC register*, report submitted to the General Medical Council, London

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

Brown, A, 2005, *Interviewer variability in oral proficiency interviews*, Peter Lang, Frankfurt

Brown, A, 2006, 'An examination of the rating process in the revised IELTS Speaking Test', *IELTS Research Reports, Volume 6*, IELTS Australia, Canberra and British Council, London, pp 1-30

Canale, M and Swain, M, 1980, 'Theoretical bases of communicative approaches to second language teaching and testing', *Applied Linguistics,* vol 1, pp 1-57

Chun, D, 2002, *Discourse intonation in L2: From theory and research to practice*, John Benjamins, Amsterdam

Council of Europe, 2001, *Common European Framework of Reference for languages: Learning, teaching, assessment*, Cambridge University Press, Cambridge

Creswell, JW and Plano-Clark, V, 2011, *Designing and conducting mixed-methods research*, 2nd ed, Sage, Thousand Oaks, CA

Crowther, D, Trofimovich, P, Isaacs, T and Saito, K, 2015, 'Does speaking task affect second language comprehensibility?' *Modern Language Journal,* vol 99, advance online access from doi:10.1111/modl.12185

Crowther, D, Trofimovich, P, Saito, K and Isaacs, T, 2014, 'Second language comprehensibility revisited: Investigating the effects of learner background', *TESOL Quarterly*, advance online access from [http://onlinelibrary.wiley.com/doi/10.1002/tesq.203/abstract]

Derwing, TM and Munro, MJ, 2009, 'Putting accent in its place: Rethinking obstacles to communication', *Language Teaching,* vol 42, pp 1-15

Derwing, TM, Thomson, RI and Munro, MJ, 2006, 'English pronunciation and fluency development in Mandarin and Slavic speakers', *System,* vol 34, pp 183-193

DeVelle, S, 2008, 'The revised IELTS pronunciation scale', *Research Notes,* vol 34, pp 36-38

Ejzenberg, R, 2000, 'The juggling act of oral fluency: A psycho-sociolinguistic metaphor', in *Perspectives on Fluency*, ed H Riggenbach, University of Michigan Press, Ann Arbor, MI, pp 287-313

ETS, 2011, 'TOEFL® program history', *TOEFL iBT® Research Insight* series 1, Educational Testing Service, Princeton, NJ, accessed from [http://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v6.pdf [11 November 2014]

Field, A, 2009, *Discovering statistics using SPSS*, 3rd ed, Sage, Thousand Oaks, CA

Galaczi, E, Lim, G and Khabbazbashi, N, 2012, 'Descriptor salience and clarity in rating scale development and evaluation', paper presented at *Language Testing Forum*, Bristol, UK, 16-18 November

General Medical Council, 2014, 'Strong support for checking doctors' language skills', *GMC News*, accessed from http://www.gmc-uk.org/publications/23811.asp [13 February 2015]

Hair, JF, Anderson, RE, Tatham, RL and Black, WC, 1998, *Multivariate data analysis*, 5th ed, Macmillan, New York

Harding, L, 2013, 'Pronunciation assessment', in *The Encyclopedia of Applied Linguistics*, ed CA Chapelle, Wiley-Blackwell, Hoboken, NJ

Hawkins, JA and Filipovic, L, 2012, *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*, Cambridge University Press, Cambridge

IELTS, 2007, *IELTS Handbook 2007*, accessed from https://www.ielts.org/pdf/IELTS_Handbook.pdf [12 February 2015]

IELTS, 2012, *IELTS Guide for Teachers*, accessed from http://www.ielts.org/PDF/IELTS_Guide_For_Teachers_BritishEnglish_Web.pdf [12 February 2015]

IELTS, 2014, *IELTS researchers: Guidelines for applying*, accessed from http://www.ielts.org/researchers/grants_and_awards/guidelines_for_applying.aspx [11 November 2014]

Isaacs, T, 2008, 'Towards defining a valid assessment criterion of pronunciation proficiency in non-native English speaking graduate students', *Canadian Modern Language Review,* vol 64, pp 555-580

Isaacs, T, 2013, 'Phonology: Mixed methods', in *The Encyclopedia of Applied Linguistics*, ed CA Chapelle, Wiley-Blackwell, Hoboken, NJ, pp 4427-4434

Isaacs, T, 2014, 'Assessing pronunciation', in *The Companion to Language Assessment*, ed AJ Kunnan, Wiley-Blackwell, Hoboken, NJ, pp140-155

Isaacs, T, accepted, 'Assessing speaking', in *Handbook of second language assessment*, eds D Tsagari and J Banerjee, DeGruyter Mouton, Berlin

Isaacs, T, submitted, 'Shifting sands in pronunciation teaching and assessment research and practice,' *Language Assessment Quarterly*

Isaacs, T, Foote, JA and Trofimovich, P, 2013, 'Drawing on teachers' perceptions to adapt and refine a pedagogically-oriented comprehensibility scale for use on university campuses', paper presented at the *Pronunciation in Second Language Learning and Teaching Conference*, Ames, IA, USA, September 20-21

Isaacs, T and Thomson, RI, 2013, 'Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions', *Language Assessment Quarterly,* vol 10, pp 135-159

Isaacs, T and Trofimovich, P, 2012, '"Deconstructing" comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings', *Studies in Second Language Acquisition,* vol 34, pp 475-505

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

Levis, JM, 2005, Changing contexts and shifting paradigms in pronunciation teaching, *TESOL Quarterly,* vol 39, pp 369-377

Levis, JM, 2006, 'Pronunciation and the assessment of spoken language' in *Spoken English, TESOL and Applied Linguistics: Challenges for Theory and Practice*, ed R Hughes, Palgrave Macmillan, New York, pp 245-270

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

Munro, MJ, 1998, 'The effects of noise on the intelligibility of foreign-accented speech', *Studies in Second Language Acquisition,* vol 20, pp 139-154.

Munro, MJ and Derwing, TM, 1999, 'Foreign accent, comprehensibility, and intelligibility in the speech of second language learners', *Language Learning,* vol 49, pp 285-310

Piske, T, MacKay, IRA and Flege, JE, 2001, 'Factors affecting degree of foreign accent in an L2: A review', *Journal of Phonetics,* vol 29, pp 191-215

Saito, K, Trofimovich, P and Isaacs, T, in press, 2016, 'Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study', *Applied Linguistics*

Saito, K, Trofimovich, P and Isaacs, T, 2015, 'Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels', *Applied Psycholinguistics*, advance online access from doi:10.1017/S0142716414000502

Segalowitz, N, 2010, *The cognitive bases of second language fluency*, Routledge, New York

Setter, J, 2008, 'Communicative patterns of intonation in L2 English teaching and learning: The impact of discourse approaches' in *English Pronunciation Models: A Changing Scene*, ed K Dziubalska-Kolaczyk and J Przedlacka, Peter Lang, Bern, pp 367-389

Trofimovich, P and Isaacs, T, 2012, 'Disentangling accent from comprehensibility', *Bilingualism: Language and Cognition* vol 15, pp 905-916

UK government website, 2014, *Tier 4 (General) student visa*, accessed from https://www.gov.uk/tier-4-general-visa/knowledge-of-english [11 November 2014]

Weir, CJ, Vidaković, I and Galaczi, E, 2013, *Measured constructs: A history of Cambridge English language examinations 1913–2012*, Cambridge University Press, Cambridge

Winke, P and Gass, S, 2013, 'The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation', *TESOL Quarterly,* vol 47, pp 762-789

Winke, P, Gass, S and Myford, C, 2013, 'Raters' L2 background as a potential source of bias in rating oral performance', *Language Testing,* vol 30, pp 231-252

Yates, L, Zielinski, E and Pryor, E, 2011, 'The assessment of pronunciation and the new IELTS Pronunciation Scale' in *IELTS Research Reports, Volume 12*, IDP: IELTS Australia, Canberra and British Council, London, pp 23-68

Zielinski, BW, 2008, 'The listener: No longer the silent partner in reduced intelligibility', *System,* vol 36, pp 69-84

APPENDIX 1: A DESCRIPTION OF THE 18 RESEARCHER-CODED MEASURES USED IN THE PRELIMINARY STUDY

Source: Isaacs and Trofimovich (2012) and Trofimovich and Isaacs (2012)

1 *Segmental error ratio*: The number of segmental (vowel, consonant) substitutions divided by the total number of segments articulated

2 *Syllable structure error ratio*: The number of segmental epenthesis (insertion) and elision (deletion) errors divided by the total number of syllables articulated

3 *Word stress error ratio*: The number of word stress errors in polysyllabic words (i.e., misplaced or absent primary stress) divided by the total number of polysyllabic words articulated

4 *Vowel reduction ratio*: The number of correctly reduced syllables over the number of obligatory vowel reduction contexts in both polysyllabic words and function words (as a measure of English stress-timed rhythm)

5 *Pitch contour*: The number of correct pitch patterns produced at the end of phrases (i.e., syntactic boundaries) over the total number of phrases where pitch patterns are expected

6 *Filled pauses*: The total number of nonlexical pauses (e.g., um, uh)

7 *Unfilled pauses*: The total number of silent pauses (≥400 ms)

8 *Pause error ratio*: The number of inappropriately produced filled and unfilled pauses (i.e., within clauses) divided by the total number filled and unfilled of pauses

9 *Repetition and self-correction ratio*: The number of immediately repeated and self-corrected words over the total number of words produced

10 *Pruned syllables per second*: The number of syllables produced excluding dysfluencies (e.g., filled pauses, repetitions, self-corrections, false starts) divided by speech sample duration

11 *Mean length of run*: The mean number of syllables produced between two adjacent filled or unfilled pauses (≥ 400 ms)

12 *Grammatical accuracy*: The number of words with at least one morphosyntactic error divided by the total word count

13 *Lexical error ratio*: The number of incorrectly used lexical expressions over the total number of words produced

14 *Token frequency*: The total number of words produced

15 *Type frequency*: The total number of unique words produced

16 *Story cohesion*: The number of adverbials used as cohesive devices (e.g., *suddenly*, *but*, *hopefully*)

17 *Story breadth*: The number of distinct propositions or storytelling elements produced (e.g., setting, initiating event, reaction)

18 *Number of story categories*: The number of different proposition categories produced

Note. Measures not already expressed as a ratio were normalised by dividing by the total duration of the analysed L2 speech sample (range: 23-36 s)

APPENDIX 2: BACKGROUND QUESTIONNAIRE

The purpose of this questionnaire is to gather information about your background as a language learner, teacher, and rater. Please answer as completely as you can.

1. Birthplace (city, country): _____ 2. Age:_____

3. Is your hearing normal as far as you know? ☐ yes ☐ no

4. First language(s) from birth: _____

5. Mother's first language: _____ 6. Father's first language: _____

7. If you were ever schooled in a language other than English as the primary medium of instruction, please specify which language in the table below. If English was the predominant language throughout your schooling, please skip to the next question.

Educational level	Language of instruction (if not English)
Primary	
Secondary	
Undergraduate	
Graduate	

8. Which languages can you speak other than English (if any)? _____

9. Of the languages you listed above, which would you say you are proficient in?
_____

10. If you have you lived outside of the UK or your country of birth for 6 months or more, please complete the following table.

Country you lived in	Time you spent there	Did you teach English while abroad?
	___ years ___months	☐ yes ☐ no
	___ years ___months	☐ yes ☐ no
	___ years ___months	☐ yes ☐ no

11. If you had exposure listening to and understanding the English language accents of any particular groups of second language speakers as part of your personal or professional connections, please specify the language(s) and reason for this increased familiarity below.

Language	Reason (e.g., family)

12. Approximately what percent of the time do you speak English (as opposed to other languages) in your daily life?

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

13. Approximately what percent of the time do you listen to the English language media (as opposed to the media in other languages)?

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

14. How many years of ESL/EFL teaching experience do you have? ____ years

15. Please indicate which university degrees and/or English teaching qualifications you have? Where appropriate, please specify your university major or programme of study (e.g., applied linguistics). You may check [√] more than one answer.

- ☐ PGCE in _____
- ☐ Diploma in _____
- ☐ Bachelor's in _____
- ☐ Master's in _____
- ☐ PhD in _____
- ☐ EdD in _____
- ☐ CELTA
- ☐ DELTA
- ☐ Trinity CertTESOL
- ☐ Trinity LTCL DipTESOL
- ☐ Other (please specify) _____

16. If you ever received pronunciation training in English or another language or have taken a phonetics/phonology course, please indicate the nature of your course/training in the table below.

Name of the pronunciation course	Additional details

17. When did you qualify as an IELTS examiner? _____ (year)

18. When did you complete your last IELTS recertification (if applicable)? _____ (year)

19. Please mark an 'X' on the below lines (scales) to approximate how comfortable you feel providing assessments on the following IELTS speaking subscales, in terms of your ability to make level distinctions.

IELTS speaking subscales	**Not comfortable at all**   **Very comfortable**
Fluency and coherence	
Lexical resource	
Grammatical range and accuracy	
Pronunciation	

20. The IELTS pronunciation scale was recently expanded from a 4-level to a 9-level scale. In which of the following ways have you received training/support on the use of this new pronunciation scale?

Face-to-face standardization (group setting)	☐ yes	☐ no
Self-access standardization (individual)	☐ yes	☐ no
Additional IELTS documentation	☐ yes	☐ no

21. Please rate how comfortable you feel rating the following terms or concepts that appear in the IELTS pronunciation subscale.

IELTS scale descriptors	**Not comfortable at all**   **Very comfortable**
Phonological features	
Connected speech	
Accent	
Intelligibility	
Rhythm	
Stress	
Intonation	
Chunking	
Stress-timing	
Speech rate	
Phonemes	

APPENDIX 3: PRE-RATING DISCUSSION GUIDELINES FOR FOCUS GROUP

1. What has been your experience rating using the 4-point (former) vs. the 9-point (current) IELTS pronunciation scale?
   - Do you prefer to rate using the longer or shorter scale?
   - To what extent do you feel that the training that you received on the 9-point IELTS pronunciation scale adequately prepared you for operational assessments?
2. How do you find the terminology that is used in the IELTS pronunciation scale?
   - Are you overall familiar with the terms?
   - Are there places where you feel that the descriptors could be clarified/improved?
   - In your view, is the clarity of the IELTS pronunciation scale descriptors on par with those of the other IELTS speaking subscales?
3. Are there particular levels that you have difficulty distinguishing between in terms of the IELTS pronunciation scale?
   - What strategy do you use to cope with band descriptors 3, 5, and 7 that state that the test-taker's performance reflects 'all of the positive features of band X and some, but not all, of the positive features of band X?'
   - Which pronunciation criteria tend to make someone a 7 and not a 6 for you? (a crucial distinction for university entrance purposes)
   - Which pronunciation criteria are most important for you in making your judgments? Are these features specifically described in the scale?
4. The pronunciation criterion, as stated in the 2007 *IELTS Handbook*, refers to 'the ability to produce comprehensible speech to fulfil the Speaking test requirements'. Does this coincide with your understanding of the pronunciation criterion?
   - How do you interpret 'comprehensible speech?'
   - 'Accent' is explicitly referred to in the scale. What role does accent play in your assessments?
5. Do you have any other comments about the IELTS pronunciation scale or rating experiences that you'd like to share before we get started with the ratings?

APPENDIX 4: INSTRUCTIONS ON RATING PROCEDURE

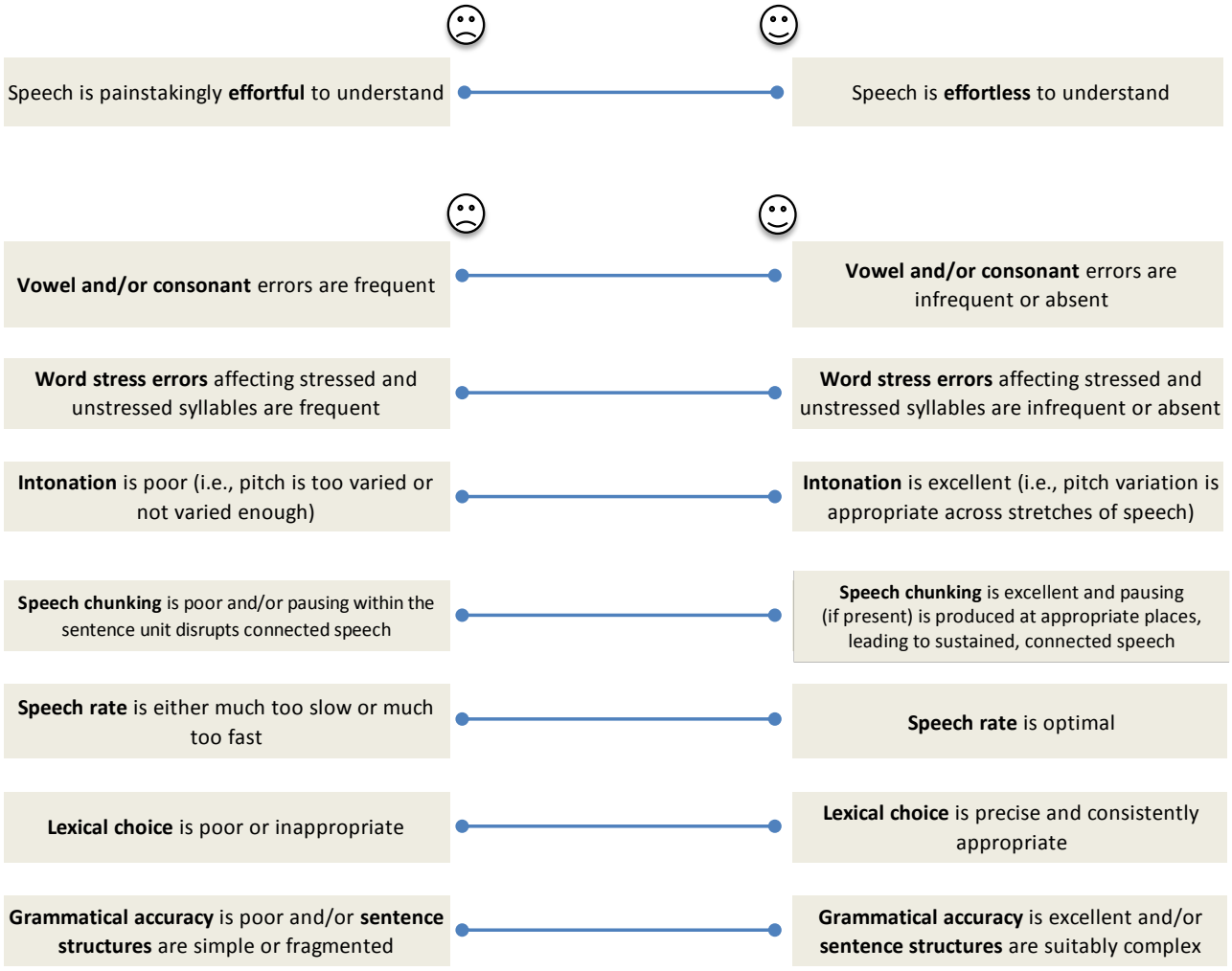
In this session, you will listen to test-takers from different IELTS test centres around the world performing the IELTS long-turn speaking task (task 2). The speech samples are variable in terms of their recording quality, with some test centres clearly with access to better recording equipment and less background noise than others.

Your task is simple. First, you will listen to each test-taker's performance on the IELTS long-turn task (task 2). You will rate the speech using the IELTS Speaking subscales. Please consult the IELTS Speaking band descriptors on the separate sheet. Please select the rating that you will assign for each subscale by circling the appropriate level on the rating sheet.

IELTS Speaking band descriptors			
Fluency & Coherence	Lexical Resource	Grammatical Range & Accuracy	Pronunciation
9	9	9	9
8	8	8	8
7	7	7	7
6	6	6	6
5	5	5	5
4	4	4	4
3	3	3	3
2	2	2	2
1	1	1	1

NOTE: It is possible that the sound quality of the recording is so poor and the background noise is so distracting that you simply cannot provide a reasonable assessment of the speech using any or all of the scales. If this is the case, please indicate this by checking 'speech is unassessble' in the little box at the top of the scoring for that test-taker, and you can skip rating that test-taker. However, please only choose this option as a last resort. If you can possibly rate all or part of the speech using the rating scales provided, please do so.

Comprehensibility:



Note. Whereas the IELTS examiners received the complete set of instructions shown here prior to conducting their ratings, the Canadian EAP teachers only received instructions on the semantic differential scales shown in the second part of this instrument. That is, the first part of the instrument featuring the IELTS band descriptors was omitted in the version devised for the Canadian EAP teachers. In addition, the paragraph explaining the 'unassessable' option was only provided to the IELTS examiners due to the potential difficulty dealing with recording quality for the IELTS files.

APPENDIX 5: DEFINITIONS FOR THE CONSTRUCTS OPERATIONALISED IN THE SEMANTIC DIFFERENTIAL SCALES

You are going to rate speech samples on several different aspects of speech. To help you with these ratings, we have included some basic definitions of the terms we are using for our rating scales.

Word	Explanation
Comprehensibility	This term refers to how much effort it takes to understand what someone is saying. If you can understand with ease, then you would consider that person highly comprehensible. However, if you struggle and must listen very carefully, or in fact cannot understand what is being said at all, then you would consider that person to have low comprehensibility.
Vowel and consonant errors	This refers to errors in individual sounds. For example, perhaps somebody says '*silly*' but you hear an 'r' sound instead of an 'l' sound. This would be a consonant error. If you hear someone say '*list*' but you hear an 'ee' sound rather than an '/ɪ/' sound (as in 'bit'), that is a vowel error. You may also hear sounds missing from words, or extra sounds added to words. These are also consonant and vowel errors.
Word stress	When an English word has more than one syllable, one of the syllables will be a little bit louder and longer than the others. For example, if you say the word '*computer*,' you may notice that the second syllable has more stress (comPUter). If you hear stress being placed on the wrong syllable, or you hear equal stress on all of the syllables in a word, then there are word stress errors.
Intonation	Intonation can be thought of as the melody of English. It is the natural pitch changes that occur when we speak. For example, you may notice that when you ask a question with a yes/no answer, your pitch goes up at the end of the question. If someone sounds 'flat' when they speak, it is likely because their intonation is not following English intonation patterns.
Speech chunking	When speaking, people naturally break speech into chunks. For example, when someone says 'how are you', it is said as one smooth chunk without any pausing. If pauses come in unnatural places, then there are problems with speech chunking.
Speech rate	Speech rate is simply how quickly or slowly someone speaks. Speaking very quickly can make speech harder to follow, but speaking too slowly can as well. A good speech rate should sound natural and be comfortable to listen to.
Lexical richness	Lexical richness refers to the vocabulary words a person uses. If people use very simple words then their speech lacks lexical richness. If incorrect or inappropriate words are used, this is also poor lexical richness. A person who demonstrates lexical richness will be comfortable with idiomatic and natural uses of English vocabulary and will have the vocabulary needed to discuss the topic he/she is speaking about.
Grammatical accuracy and sentence structure	This refers to the number of grammatical errors that the speaker makes, including word order, and/or the simplicity or complexity of the clauses or sentences that the speaker attempts.

APPENDIX 6: INSTRUMENT FOR RECORDING RATINGS FOR EACH SPEECH SAMPLE

IELTS Speaking band descriptors			
Fluency & Coherence	Lexical Resource	Grammatical Range & Accuracy	Pronunciation
9	9	9	9
8	8	8	8
7	7	7	7
6	6	6	6
5	5	5	5
4	4	4	4
3	3	3	3
2	2	2	2
1	1	1	1

Note. The IELTS examiners recorded their ratings while consulting the IELTS Speaking band descriptors (official version) for the scale on the left hand side of the page and the semantic differential scales (Appendix 4) and accompanying definitions (Appendix 3) to complete the scale on the right hand side of the page.

APPENDIX 7: POST-RATING SUMMARY OF IMPRESSIONS

Please list any additional criteria that do not already appear on this list at the top of the chart below and then provide your rankings.

	Criteria used for rating using the IELTS pronunciation scale	**Your rank** (1 = most important, 2 = 2nd most important, **N/A = not relevant**)
Additional criterion ➔		
Additional criterion ➔		
Additional criterion ➔		
	Vowel and consonant errors	
	Word stress errors	
	Intonation	
	Speech chunking and pausing	
	Speech rate	
	Lexical Choice	
	Grammatical accuracy and sentence structure	

APPENDIX 8: POST-RATING DISCUSSION GUIDELINES FOR FOCUS GROUP

1. Referring back to the question about your rank ordering of criteria in the questionnaire, could you explain which criteria you chose as the most important influences on your judgments in today's data collection session?

   - Is this different from your experience conducting IELTS pronunciation ratings in general?

2. Are there any criteria not represented in the IELTS pronunciation scale that influenced your judgments? (These may or may not have been featured in the semantic differential scales)

3. Did you have any particular difficulties using the IELTS pronunciation scale?

   - interpreting what is meant by the scale criteria?
   - distinguishing between adjacent levels of the scale?
   - other?

4. How did you distinguish between IELTS pronunciation scale bands 6 and 7?

   - Do you feel comfortable with the ratings you assigned for the discrete scales and overall comprehensibility?

5. In terms of using the discrete (semantic differential) scales in today's data collection session, were there some measures that proved more difficult to assign scores for than others?

6. We asked you to provide ratings for both 'lexical choice' and 'grammatical accuracy and sentence structure', which do not fall under the remit of the IELTS pronunciation scale. How did these relate to the other discrete criteria that you assessed?

7. Did the quality of the recordings affect your ratings?

8. We used the IELTS long-turn task as part of this study. Do you have any comments about use of the IELTS pronunciation scale with the other two speaking tasks (including when there is more interaction with the IELTS examiner)?

9. Do you have any other comments about the IELTS pronunciation scale or about today's data collection session?