IELTS™

The assessment of pronunciation and the new IELTS Pronunciation scale

Authors
Lynda Yates
Macquarie University, Australia

Beth Zielinski
Macquarie University, Australia

Elizabeth Pryor
The University of Melbourne, Australia

Grant awarded Round 14, 2008

The revised Pronunciation scale of the IELTS Speaking Test became operational in August 2008 and expanded the four bands to nine bands, in line with the three other analytic scales. This study explores examiners' perceptions, experiences and behaviour as they use the new scales with speakers from two different language backgrounds at the crucial Pronunciation band levels of 5, 6 and 7.

Click here to read the Introduction to this volume which includes an appraisal of this research, its context and impact.

ABSTRACT

Using a mixed method approach, this study explores how examiners view the revised IELTS Pronunciation scale in general, and how they used the scale to award Pronunciation scores to candidates at IELTS-assigned scores of Bands 5, 6 and 7. In general, the examiners preferred the revised Pronunciation scale to the previous one, and they were largely positive about how easy the descriptors and increased number of band levels were to use. They also reported feeling confident about assessing the different features of pronunciation covered in the Pronunciation scale descriptors, and most confident about making global judgements of intelligibility and listener effort, which were the features they considered to be the most important when awarding a Pronunciation score. However, when using the scale to award Pronunciation scores, there was considerable variation between the scores awarded and the features identified as contributing to the assessment of the candidates' pronunciation. The distinction between Bands 6 and 7 seemed to be particularly problematic, and there was a tendency to award a 6 to Band 7 speakers. The examiners expressed concerns in relation to the specificity of the descriptors at Bands 3, 5 and 7 and the overlap between the Pronunciation scale and the Fluency and Coherence scale.

These findings suggest that further revision of the descriptors and documentation related to the scale may be useful and that increased attention to pronunciation in the selection, training and professional development of examiners.

AUTHOR BIODATA

LYNDA YATES

Lynda Yates is Associate Professor in Linguistics in the Department of Linguistics, Macquarie University. She has taught adults in a wide range of industrial and educational settings and has a keen interest in teacher professional development. Her research interests centre on language learning by adults and include intercultural pragmatics, workplace communication, issues for immigrants and the teaching and learning of spoken English and pronunciation. Lynda and Beth Zielinski have recently published an introductory volume on teaching pronunciation to adults.

BETH ZIELINSKI

Beth Zielinski is a Postdoctoral Research Fellow in the Department of Linguistics, Macquarie University. Her research interests are in the area of pronunciation and intelligibility, and her doctoral thesis investigated the features of pronunciation that have an impact on intelligibility in speakers of English as a second language. She has conducted pronunciation classes for international university students, private consultations for corporate clients, and professional development sessions for teachers, as well as publishing and lecturing in the area. She has recently published an introductory volume on teaching pronunciation to adults with Lynda Yates.

ELIZABETH PRYOR

Liz Pryor has previously taught ESL in Australia and EFL at the British Council in Singapore where she certified as an IELTS examiner. She has also worked in teacher education and communication skills training. She is currently working in the field of healthcare communication, and is completing an MA in this area at the University of Melbourne.

IELTS RESEARCH REPORTS, VOLUME 12, 2011

Published by: IDP: IELTS Australia and British Council
Editor: Jenny Osborne, IDP: IELTS Australia
Editorial consultant: Petronella McGovern, IDP: IELTS Australia
Editorial assistance: Judith Fairbairn, British Council
Acknowledgements: Dr Lynda Taylor, University of Cambridge ESOL Examinations

IDP: IELTS Australia Pty Limited
ABN 84 008 664 766
Level 8, 535 Bourke St
Melbourne VIC 3000, Australia
Tel +61 3 9612 4400
Email ielts.communications@idp.com
Web www.ielts.org
© IDP: IELTS Australia Pty Limited 2011

British Council
Bridgewater House
58 Whitworth St
Manchester, M1 6BB, United Kingdom
Tel +44 161 957 7755
Email ielts@britishcouncil.org
Web www.ielts.org
© British Council 2011

This publication is copyright. Apart from any fair dealing for the purposes of: private study, research, criticism or review, as permitted under the Copyright Act, no part may be reproduced or copied in any form or by any means (graphic, electronic or mechanical, including recording, taping or information retrieval systems) by any process without the written permission of the publishers. Enquiries should be made to the publisher. The research and opinions expressed in this volume are of individual researchers and do not represent the views of IDP: IELTS Australia Pty Limited. The publishers do not accept responsibility for any of the claims made in the research.

National Library of Australia, cataloguing-in-publication data, 2011 edition, IELTS Research Reports 2011 Volume 12
ISBN 978-0-9775875-8-2

CONTENTS

1 INTRODUCTION

The Pronunciation scale of the IELTS Speaking Test has been revised and it now has nine bands, in line with the three other analytic scales. In addition, there has been a shift from judgements on global aspects of pronunciation, such as intelligibility, to the assessment of a number of specific phonological features. Thus, the IELTS Pronunciation descriptors include both attention to accuracy in segmental and prosodic features, as well as evaluations of the more global constructs of intelligibility, listener effort and accentedness. The interrelationships between these different aspects of pronunciation are complex and little is known about how examiners integrate the judgements of these parameters to arrive at a single, summary Pronunciation score for a particular sample.

After extensive re-training and re-certification in centres around the world, the revised Pronunciation scale became operational in August 2008. The aim of this study is to explore examiners' perceptions, experiences and behaviour as they use the new scales with speakers from two different language backgrounds at the crucial Pronunciation band levels of 5, 6 and 7.

2 ISSUES RELATED TO PRONUNCIATION AND ITS ASSESSMENT

Pronunciation is a crucial 'first level hurdle' for learners to master because if their performance cannot be understood, it cannot be rated on any other scale (Iwashita, Brown, McNamara & O'Hagan 2008, p 44). Therefore, it is a vital component of proficiency in spoken English, yet it does not always receive the attention it deserves in either the teaching or the testing literature or in teacher training. While there has been some renewed interest in the field of pronunciation learning and teaching in recent years (for example, a special issue of *TESOL Quarterly*, 2005 edited by Levis; *Prospect*, 2006 edited by Yates; Derwing & Munro 2005; Levis 2006; Hansen Edwards & Zampini 2008), there is still little published work on pronunciation in spoken assessment. Furthermore, the precise identification of pronunciation problems can be difficult even for experienced listeners. Schmid and Yeni-Komshian (1999), for example, found that native speaker listeners had increased difficulty detecting mispronunciations at the phonemic level as accentedness increased, and Derwing and Rossiter (2003) found similar issues among the experienced listeners in their study. Research has indicated that many teachers lack training and confidence in their expertise in pronunciation learning and teaching (Levis 2006, Macdonald 2002). This suggests that the skills needed by examiners to assess this area of language use may be in relatively short supply as it is likely to be an area that they find challenging. It also raises questions of which features examiners are able to identify as problematic and how they relate these to the new, more differentiated descriptors.

Studies that have addressed the assessment of spoken English suggest that examiners may not feel as comfortable judging pronunciation as they do other aspects of a speaker's performance. In their pilot study of assessment processes in the Certificate in Advanced English (CAE) Speaking test, for example, Hubbard, Gilbert and Pidcock (2006) found that, of the four CAE analytical criteria, examiners commented the least on pronunciation in their real time verbal protocol analysis, and that the comments they did make were largely either general in nature or related to individual sounds. Hubbard et al suggest that this may have been related to the examiners' tendency to make their decision about pronunciation in the first part of the test, and not comment on it further in the latter parts. However, the general nature of their comments might also indicate that they had difficulty identifying pronunciation features other than individual sounds. Brown (2006) also observed a similar trend in examiner comments related to pronunciation in her study of the previous version of the IELTS Speaking Test, which included the four-band version of the Pronunciation scale.

The way in which examiners make use of a scale can also vary, so it is imperative to investigate how the revised Pronunciation scale is being used to ensure that it is used consistently as envisaged by test developers (Orr 2002). In Brown's study on the rating processes (2007), she found that examiners oriented in general to scale descriptors but she also noted variability in their scoring behaviour. She explained this variability in terms of a tendency to interpret criteria differently, make use of criteria not in the descriptors and to give a differential weighting to the different criteria. Orr (2002) noted a similar variation, and in view of the inherent difficulty of assessing spoken performance and of the use of scales in particular, he concluded that there is a need to focus on both process and product in rater training. He also noted the importance of understanding exactly what scores on speaking tests reflect, and was hopeful in this regard about the potential of the scales being developed at that time for the IELTS Speaking Test (p 153).

The revised IELTS Speaking Test became operational in 2001. It introduced a tightly scripted interview format and moved away from a single holistic band scale for assessment to four analytical assessment scales focussing on: Grammatical Range and Accuracy; Lexical Resource; Fluency and Coherence; and Pronunciation (Brown & Taylor 2006). All of the scales made use of a nine-point scale except for Pronunciation, which used a four-point scale (2, 4, 6, 8). In their study on views and experiences using the revised test, Brown and Taylor (2006) reported a largely positive response from the 269 examiners surveyed, but found that the Pronunciation scale was consistently reported as less easy to interpret and less capable of discriminating between levels than the other three scales. Over half of the examiners identified Pronunciation as the scale about which they felt the least confident (p 3). There was also some indication that the provision of only four bands might encourage the awarding of a Band 6 score by default (Brown 2006). A revised Pronunciation scale was therefore developed in two major phases in 2007 to cover all nine points on the scale. Results of trialling in the first phase were generally positive, but also highlighted issues with the small number of positive features available for use at lower levels. The 10 examiners who trialled in the second phase were largely positive about the new scale, showed little variation for harshness and did not appear to have any issues with the wording of the descriptors. The revised draft was therefore introduced in August 2008 following extensive examiner re-training (De Velle 2008).

The revised Pronunciation scale constitutes an explicit move away from global judgements of intelligibility towards descriptors that clearly specify 'key performance pronunciation features' which examiners are trained to identify in a candidate's performance (De Velle 2008, p 36). These features, and the extent to which they are mastered, are listed in the descriptors at Bands 2, 4, 6, 8 and 9, while the descriptors at the new Bands 3, 5 and 7 are more general in nature and invite the examiner to compare the performance against the features listed at the levels above and below. For example, to be awarded a Band 5 score, a performance needs to display 'all the positive features of Band 4 and some, but not all, of the positive features of Band 6' (IELTS 2010).

Global judgements of intelligibility, listener effort, clarity and accent are also mentioned or referred to at various band levels. There is no definition provided in the IELTS documentation for the term 'clarity', but the terms 'intelligibility', 'listener effort' and 'accent' appear to correspond to Derwing and Munro's (2005) now widely accepted definitions:

- intelligibility – 'the extent to which a listener actually understands an utterance' (p 385) or is able to decode a message
- comprehensibility – 'a listener's perception of how difficult it is to understand an utterance' (p 385)
- accentedness – 'a listener's perception of how a speaker's accent is different from that of the L1 community' (p 385).

Work by Munro, Derwing and colleagues (see for example, Derwing & Munro 1997, Munro & Derwing 1995a, 1995b, Munro, Derwing & Morton 2006) has investigated the impact of a range of features on assessments of these three dimensions of pronunciation and has identified them as related but independent. Intelligibility scores as measured by a transcription task correlated more strongly with comprehensibility than accentedness. Some L2 (second language) speakers can be understood, that is listeners are able to understand the content of what they are saying, but this understanding may take considerable effort. One of Derwing and Munro's most robust findings is that a speaker's degree of accentedness is not a good indicator of their intelligibility score, and therefore an accent *per se* is not necessarily a problem.

A distinction between intelligibility, listener effort and accent is reflected in the Pronunciation descriptors, but the relationship between these global judgements and the more specific key features is not always clear. Previous studies have highlighted the role of different pronunciation features in these judgements. Word stress (Field 2005) and primary stress (Hahn 2004) have been found to play an important role in a listener's understanding of an utterance. The appropriate production of intonation units has been found to be more characteristic of higher-level learners (Iwashita, Brown, McNamara & O'Hagan 2008). Segmental deviations have been found to significantly impact on accent ratings for both ESL students (Munro & Derwing 2001) and learners of Swedish as a second language (Boyd 2003). Segments in word initial position (Schmid & Yeni-Komshian 1999; Bent, Bradlow & Smith 2007) and strong syllables (Zielinski 2008) have been found critical to judgements of intelligibility, as has vowel production accuracy as opposed to overall consonant accuracy (Bent, Bradlow & Smith 2007).

The small number of studies that have investigated examiner behaviour and the relationship between test performances and the score awarded suggest that examiners can arrive at similar scores for different reasons and award different scores, although they assess the same performance in a similar way (Brown 2006, Orr 2002). There is a need for more such studies to illuminate how examiners use scales and arrive at the scores they award.

In this study we focus on the use of the revised Pronunciation scale, in particular on band levels 5, 6 and 7 for two major reasons. First, the revised scale involved the addition of the 'in-between' bands of 5 and 7 in a bid to give examiners greater flexibility in awarding scores than they had had using the previous four-point scale. Secondly, differentiating between these bands is crucial in an Australian context where the attainment of overall band levels between 5.5 and 7 can often be high stakes for test-takers. In Australia, an overall band level of 6.0 (or in some cases 6.5) is required for entry to undergraduate study. Entry to some professional courses such as teaching and nursing often requires a score of at least 7.0 on the spoken component. Moreover, in Australia, additional points can be gained for permanent resident applications if an overall score of 7.0 is obtained. This level of IELTS is also required for entry to some professions, so that failure to gain a 7 on just one of the four scales can make a crucial difference to whether a candidate can practise their profession in Australia. However, there has been considerable media comment recently on the adequacy of the language competence of graduates who have succeeded in obtaining permanent resident status but whose spoken competence in particular is perceived as inadequate for Australian workplaces (Birrell & Healy 2008).

This study takes a mixed method approach to allow investigation of both how examiners view the revised Pronunciation scale in general and how they use it to award scores at these crucial band levels.

3 METHODOLOGY

3.1 Research questions

The design of this study allowed for the exploration of how examiners view the revised IELTS Pronunciation scale in general, and how they used the scale to award a Pronunciation score to candidates at band levels 5, 6 and 7. The mixed method nature of the study allowed for the analysis of both quantitative and qualitative data collected from a number of different sources to address the following research questions.

1. In general, how do examiners view the revised IELTS Pronunciation scale?
    a) How easy do they find the descriptors and increased number of bands to use?
    b) How confident do they feel about judging a candidate's use of the different features of pronunciation covered in the descriptors?
    c) Which features of pronunciation do they think are most important when awarding a Pronunciation score?
2. When using the revised IELTS Pronunciation scale to award a Pronunciation score to candidates at band levels 5, 6 and 7:
    a) How easy do examiners find it to distinguish between the different band levels?
    b) Which features of pronunciation do examiners take into consideration?
    c) What problems do examiners report regarding the use of the scale?

3.2 Data collection

There were three phases of data collection.

1. *Phase 1, Pre-rating:* The online Questionnaire A elicited background details and experiences with, and attitudes towards, the revised Pronunciation scale in general from 27 examiners.
2. *Phase 2, Rating*: All but one of the same group of examiners (n=26):
    - used the revised IELTS Pronunciation scale to score 12 sample performances of Part 3 of the Speaking Test (Questionnaire B)
    - completed Questionnaire C which invited them to reflect on how they had used the scale to award a Pronunciation score to sample performances.
3. *Phase 3, Verbal protocol:* A different group of examiners (n=6) each scored four of the 12 sample performances used in Phase 2 and summarised their reasons for the scores they awarded. For each of the samples, they also used a stimulated verbal protocol procedure to reflect on the features that contributed to their assessment of the candidate's pronunciation.

Copies of questionnaires A, B and C used in Phases 1 and 2 are provided in Appendix 1.

3.3 Participants

3.3.1 Examiners

All examiners were currently certified IELTS examiners from a single centre, whose examining experience ranged from newly trained (two months experience) to very experienced (13 years). All held qualifications that met the requirements for IELTS examiners, that is, an undergraduate degree (or equivalent), a relevant TESOL qualification, and at least three years relevant teaching experience (IELTS, 2010). Some had a Certificate in English Language Teaching to Adults (CELTA) qualification or an undergraduate or postgraduate TESOL qualification, and some had both.

A new group of examiners was recruited for Phase 3 of the study to ensure that they had not rated the samples previously. As shown in Table 1, they had similar TESOL qualifications as those recruited for the first two phases and had been teaching and examining for a similar length of time.

Participation in the study was voluntary, and paid at standard hourly IELTS examiner rates.

	Phase 1 and 2 examiners (n=27[a])	**Phase 3 examiners (n=6)**
Teaching experience	3 – 30 years (*M* = 14.0)	7 – 20 years (*M* = 15.0)
Experience as an IELTS examiner	newly trained – 13 years (*M* = 3.3)	newly trained – 10 years (*M* = 3.5)
CELTA Qualification	13 (48.2%)	3 (50.0%)
Undergraduate or postgraduate TESOL qualification	20 (74.1%)	4 (66.7%)

[a] 27 completed Questionnaire A, from which the information reported here was taken. Only 26 of these participated in Phase 2.

Table 1: Characteristics of participating IELTS examiners

3.3.2 Candidate speech samples

The 12 speech samples were provided by IELTS Australia and comprised excerpts (Part 3) from IELTS Speaking Test interviews of candidates from two language backgrounds, Punjabi and Arabic, who had been awarded Pronunciation band scores of 5, 6 and 7. There were two samples from each language group at each level. As shown in Table 2, this means there were six from each language group and four (two Punjabi and two Arabic) at each band level. Most of the samples were from male candidates; only one from each language group was female. Part 3 of the Speaking Test was chosen because it provided an extended sample of the candidate's spoken English in interaction (a discussion usually lasting four to five minutes) for the examiners to rate, and also because it has been argued to show the best correlation with marks on the full test (IELTS, 2010).

Band level	**Language backgrounds**		**Total**
	Punjabi	**Arabic**	
5	2	2	4
6	2	2	4
7	2	2	4
Total	**6**	**6**	**12**

Table 2: Speech samples: distribution of band level and language

3.4 Procedure

Phase 1 (pre-rating phase)

Questionnaire A was administered electronically and, where possible, the examiners completed and returned it several days before the predetermined date of the Phase 2 data collection session.

Phase 2 (rating phase)

The speech samples were randomised into four different orders and burned onto separate CDs for use by the examiners. The rating tasks were conducted in a section of a library facility where each examiner had access to their own individual computer or CD player to listen to the samples using headphones. A practice sample was presented before the 12 samples to be rated. The examiners used the revised Pronunciation scale to score each sample and recorded the scores in Questionnaire B. They were able to rate the samples at their own pace, and to use the recordings as they would if they were examining the candidates in the samples, that is, they could pause or replay them where necessary. Once they had awarded scores to all 12 samples, the examiners completed Questionnaire C.

Phase 3 (verbal protocol phase)

A number of studies of the rating process on oral proficiency tests have used examiners' retrospective verbal reports to focus on the decisions they make when judging a candidate's performance. Orr (2002) has researched this in terms of the examiner in the role of assessor for the Cambridge First Certificate in English (FCE) Speaking test. Brown used verbal protocol studies to provide information about the previous holistic IELTS band scales (Brown 2007) and to investigate the validity of the more recent analytic scales (Brown 2006). Recently, Hubbard et al (2006) reported positively on findings using a 'real time' verbal protocol analysis to study the Cambridge CAE Speaking test. The verbal protocol procedure used in this study draws on that used by Brown (2006).

Each examiner participating in the verbal protocol phase (VP phase) rated for pronunciation and provided verbal reports on four different samples selected from the 12 speech samples. These were selected in such a way that each VP examiner reported on samples involving both language groups and the range of IELTS-assigned band levels. Each sample was treated by two different VP examiners (see Table 3).

Each VP session took place in a quiet room with only the VP examiner and a researcher present. The samples were played through a computer with external speakers and the VP examiner paused the recording using the computer keyboard. Each VP session was recorded using a digital voice recorder and took the following format.

1. *Practice stage:* examiners practised with a sample (not included in the 12 samples) before listening to the four samples assigned to them.
2. *The rating stage:* the VP examiner was instructed to listen to the recording as they would if they were examining the candidate and to award a band score for Pronunciation. The examiner was also asked to summarise the reasons for choosing that score. The order of the samples presented to each VP examiner was randomised so that each heard samples from the different language backgrounds and IELTS-assigned band levels in a different order.
3. *The review stage:* the VP examiner was instructed to listen to the recording again and pause it to comment whenever she came across anything that contributed to her assessment of the candidate's pronunciation.

4. *The reflection stage:* the VP examiner either commented spontaneously, or if she did not, was asked for additional comments after each recording had finished.
5. *Opportunity for general comments:* after completing all four verbal reports, the VP examiner was asked for any further comments. Where necessary, the researcher also followed up on any comments the VP examiner had made during the session that needed clarification. This was the only point during the session when the researcher was engaged with the VP examiner in this way because of the potential for such discussion to affect subsequent verbal reports. Thus during the previous stages, the researchers provided minimal feedback to the VP examiners 'intended as no more than tokens of acceptance of what they said' (Lumley 2005, p 119).

In total, 24 verbal reports (six examiners by four samples) were recorded and transcribed. For one report (VP2's report on sample 7P2), the rating stage summary was not recorded but written down by the researcher due to a technical problem.

Sample[a]	VP phase examiner					
	VP 1	VP 2	VP3	VP4	VP5	VP6
5A1	X		X			
5A2		X		X		
5P1	X				X	
5P2		X				X
6A1	X				X	
6A2		X				X
6P1			X	X		
6P2			X	X		
7A1			X		X	
7A2				X		X
7P1	X				X	
7P2		X				X

[a] The labels here refer to the IELTS-assigned band level (5, 6 or 7), the language background of the candidate in the sample (A: Arabic, P: Punjabi) and the two different candidates from each language background (1 and 2).

Table 3: Samples rated by VP examiners

3.5 Summary of study design and aims

An overview of the data sources in relation to the research questions is presented in Table 4. As noted above, copies of the questionnaires used in Phases 1 and 2 are in Appendix 1.

Study phase	Data source	Purpose (research question addressed)
1 Pre-rating: n = 27	Questionnaire A	• To elicit background details of examiners • To elicit examiners' views of the revised Pronunciation scale in general (Research Question 1)
2 Rating: n = 26	Questionnaire B	• To record the Pronunciation score awarded for each sample (Research Question 2a)
	Questionnaire C	• To elicit examiners' reflections on using the revised Pronunciation scale to award Pronunciation scores to the candidates in the samples (Research Questions 2a and 2c)
3 Verbal protocol: n = 6	Rating stage	• To record the Pronunciation score awarded for each sample (Research Question 2a) • To elicit examiners' reasons for awarding a particular score (Research Question 2b)
	Review stage	• To elicit reflections on the features of pronunciation that contributed to the Pronunciation score awarded (Research Question 2b) • To elicit responses highlighting difficulties using the revised Pronunciation scale (Research Question 2c)

Table 4: Overview of data sources

3.6 Data analysis

Quantitative data from the questionnaires were analysed using SPSS to provide means and standard deviations for Likert scale responses in line with the approach taken in previous research by Derwing and Munro (see for example, Derwing & Munro 1997, Munro & Derwing 1995a). Paired-sample *t* tests were used to investigate differences between Likert scale responses at a .05 significance level.

Qualitative data from the questionnaires were coded manually by two researchers for themes related to the relevant research question. VP data were coded and analysed using NVivo 8 by one author who established coding category descriptions (see Appendix 2). A coding reliability check was performed by a second author who used the descriptions to independently code 10% of the comments selected randomly. There were very few disagreements, but these were discussed and resolved.

4 RESULTS

4.1 Research Questions 1 and 1A)

Research Question 1: In general, how do examiners view the revised IELTS Pronunciation scale?

Research Question 1a): How easy do they find the descriptors and increased number of bands to use?

Responses on Questionnaire A indicated that examiners who had used the previous version of the Pronunciation scale (n = 21) preferred the revised scale. Their comments suggested that it enabled them to be more precise and flexible in their judgements (11 and seven comments respectively), and this made it fairer for candidates (five comments).

However, the results on how easy they found in to use were less clear. Examiners were asked in Question 1 (A1) to indicate on a five-point Likert scale (1 = very easy, 5 = very hard) how easy they found it to use the descriptors of the four rating scales in the Speaking Test. Table 5 shows the means of their responses. It can be seen that the mean rating for the Pronunciation scale was 2.81 compared to 2.59, 2.41 and 2.26 for Grammatical Range & Accuracy, Fluency & Coherence, and Lexical Resource respectively, suggesting that they found it a little harder to use than the other scales. However, only the difference between the means for Pronunciation and Lexical Resource was significant (see Appendix 3).

	Fluency & Coherence	Lexical Resource	Grammatical Range & Accuracy	Pronunciation
M	2.41	2.26	2.59	2.81
SD	1.047	0.903	0.971	0.962

Table 5: Ease of use of descriptors on all scales of the Speaking Test

Table 6 shows a breakdown of examiner responses to this item (first row) and responses to Question A3 which elicited views on how easy examiners found it to use specific aspects of the Pronunciation scale. From this, we can see that the examiners tended towards the mid-point when responding to all four items.

This could be for a number of reasons. For example, a rating of 3 might indicate that the examiner found a particular aspect of using the scale neither easy nor hard, and so opted for a more neutral rating. They might also have difficulty deciding on a rating and opt for the mid-point because the degree of difficulty using different aspects of the Pronunciation scale descriptors might depend on the candidate being examined. A mid-point response to A1 could also mean that the descriptors on the Pronunciation scale were relatively easy to use, but not as easy as the other scales in the Speaking Test that were rated at the same time. This was the case for three examiners who each gave a rating of 3 for how easy they found the Pronunciation scale descriptors to use, but ratings of either 1 or 2 for how easy they found the descriptors on the other scales. The reasons they gave were as follows:

E5: *All the descriptors are relatively easy to use. The reason why pronunciation is a bit harder is because of the 'in between' bands.*

E19: *The pronunciation descriptors are relatively new compared to the others.*

E26: *The descriptors are overall succinct and user-friendly. The pronunciation descriptors I find a little vague.*

	Examiner ratings (n=27)				
	1 (very easy)	2	3	4	5 (very hard)
A1. How easy have you found it to use the Pronunciation descriptors?	2 (7.4%)	8 (29.6%)	11 (40.7%)	5 (18.5%)	1 (3.7%)
A3 (a). How easy do you find it to use the increased number of band levels?	5 (18.5%)	9 (33.3%)	9 (33.3%)	4 (14.8%)	0 (0%)
A3 (b). How easy do you find it to distinguish between the band levels	3 (11.1%)	8 (29.6%)	10 (37.0%)	5 (18.5%)	1 (3.7%)
A3 (c). How easy do you find it to understand the descriptors	3 (11.1%)	7 (25.9%)	12 (44.4%)	4 14.8%)	1 (3.7%)

Note: Percentages may not add to 100% because of rounding to one decimal place.

Table 6: Ease of use of descriptors on the revised Pronunciation scale

Some clarity can be brought to this situation through examination of those ratings at either end of the scale, that is, where examiners found an aspect of the scale easy (ratings of 1 or 2) or hard (ratings of 4 or 5). As shown in Table 6, examiners tended to be more positive than negative about how easy they found the Pronunciation scale descriptors to use.

- Almost twice as many examiners found the descriptors easy to use (n = 10) than hard to use (n = 6).
- More than three times as many examiners found the increased number of band levels easy to use (n = 14) as those who found them hard (n = 4).
- Almost twice as many examiners found it easy to distinguish between the band levels (n = 11) as those who found it hard (n = 6).
- Twice as many examiners found it easy to understand the descriptors (n = 10) as those who found it hard (n = 5).

The above findings indicate that the examiners preferred the revised Pronunciation scale to the previous one, and although responses were mixed, they tended to be more positive than negative about how easy the descriptors and increased number of band levels were to use. As will be discussed in the next section, they also reported feeling confident about assessing the different features of pronunciation covered in the Pronunciation scale descriptors.

4.2 Research Question 1B)

Research Question 1b): How confident do they feel about judging a candidate's use of the different features of pronunciation covered in the descriptors?

The examiners indicated they felt confident judging the features of pronunciation covered in the Pronunciation scale descriptors. They were asked to indicate on a five-point Likert scale how confident they felt when judging the features covered in the Pronunciation scale descriptors: sounds, rhythm, stress (word level), stress (sentence level), intonation, chunking, speech rate, intelligibility, listener strain (listener effort) and accent. As with the confidence scale used by Brown (2006), the lower the rating, the lower the degree of confidence they felt in judging the feature (1 = not very confident, 5 = very confident).

As can be seen from Table 7, which shows the means for each feature, the examiners felt quite confident in judging all of these, but reported greater relative confidence in judging the global features such as intelligibility (*M* = *4*.19), listener effort (*M* = 4.07) and accent (*M* = 3.96). Apart from speech rate (*M* = 3.96), they had slightly lower levels of confidence in judging the specific or concrete features such as rhythm (*M* = 3.52), sentence stress and intonation (*M* = 3.67), chunking and word stress (*M* = 3.74) and sounds (*M* = 3.78).

A series of paired-sample t-tests revealed that the mean for intelligibility was significantly higher than the means for all of the concrete features except speech rate, and the mean for listener effort was significantly higher than for all the concrete features except for chunking and speech rate. This was not the case for judgements of accent, however, where there was no significant difference between the means except for rhythm (see Appendix 3 for details).

	Concrete features							**Global features**		
	Sounds	Rhythm	Word stress	Sentence stress	Intonation	Chunking	Speech rate	Accent	Listener effort	Intelligibility
M	3.78	3.52	3.74	3.67	3.67	3.74	3.96	3.96	4.07	4.19
SD	0.934	0.849	1.095	1.074	1.038	1.023	0.940	0.980	1.072	1.001

Table 7: Confidence judging features of pronunciation

From the above findings, it seems that while the examiners felt confident judging all of the features covered in the Pronunciation scale descriptors, they were more confident in judging the global features (intelligibility and listener effort) than in judging most of the concrete features. As will be discussed in the next section, these were also the features they felt were most important when awarding a Pronunciation score.

4.3 Research Question 1C)

Research Question 1c): Which features of pronunciation do they think are most important when awarding a pronunciation score?

The examiners' response on Questionnaire A indicated that they thought intelligibility and listener effort were the most important features when awarding a Pronunciation score. As a follow-up to the question on how confident they felt when judging different features, examiners were asked to nominate those they felt to be the most important when awarding a Pronunciation score, and rank them if appropriate. Table 8 shows that 85.2 % of the examiners nominated intelligibility as an important feature and 21 (77.8%) ranked it either first or second in importance. Listener effort was the second most commonly nominated feature (70.4%), and 14 (51.9%) examiners ranked it either first or second in importance. Further analysis revealed that 11 examiners ranked intelligibility and listener effort as the two most important features. Of these, nine ranked intelligibility first and listener effort second.

Feature	Examiner rankings (n = 27)						Total	
	1	2	3	4	5	6	n	%
Intelligibility	16	5	1	1	0	0	**23**	**85.2%**
Listener effort	4	10	1	4	0	0	**19**	**70.4%**
Chunking	0	4	3	3	3	0	**13**	**48.1%**
Word stress	2	3	2	1	0	1	**9**	**33.3%**
Rhythm	2	3	1	1	1	0	**8**	**29.6%**
Intonation	0	0	3	2	0	2	**7**	**25.9%**
Sounds	2	0	3	0	0	0	**5**	**18.5%**
Speech rate	1	0	2	1	0	1	**5**	**18.5%**
Sentence stress	0	1	2	0	1	1	**5**	**18.5%**
Accent	0	0	1	0	0	1	**2**	**7.4%**

Table 8: Features considered most important when awarding a Pronunciation score

In summary, it seems that examiners preferred the revised IELTS Pronunciation to the previous version. Before rating the samples in this study, they were more positive than negative about how easy the descriptors were to use and felt confident about judging the different features covered in the descriptors. They felt most confident about judging the global features (intelligibility and listener effort), and considered these to be the most important features when awarding a Pronunciation score. As will be discussed below, however, when actually rating the samples in this study, the examiners did experience some difficulty in distinguishing between the band levels selected for focus.

4.4 Research Question 2A)

Research Question 2a): When using the revised IELTS Pronunciation scale to award a pronunciation score to candidates at band levels 5, 6 and 7, how easy do examiners find it to distinguish between different band levels?

The examiners seemed to find it more difficult than expected to distinguish between different band levels when awarding Pronunciation scores to the samples in this study. As shown previously in Table 6 (A3b), before rating the samples, the examiners were largely positive about how easy they found it to distinguish between band levels. However, their responses to a similar question on Questionnaire C (C1), which asked them to indicate on a similar five-point Likert scale (1 = very easy, 5 = very hard) how easy they found it to distinguish between band levels for the candidates in the samples, were not so positive.

Table 9 summarises the examiners' responses, both before and after they rated the samples. Those before rating the samples (A3b, see Table 6) are an indication of how the examiners felt in general about distinguishing between different band levels. Those provided after rating the samples (C1) indicate how they felt about using the scale to distinguish between the samples they had just rated, that is, from Arabic and Punjabi speakers at levels 5, 6 and 7.

	Examiner ratings				
	1 (very easy)	**2**	**3**	**4**	**5 (very hard)**
Before rating the samples (A3b) (n = 27)	3 (11.1%)	8 (29.6%)	10 (37.0%)	5 (18.5%)	1 (3.7%)
After rating the samples (C1) (n = 25[a])	0	4 (16.0%)	14 (56.0%)	7 (28.0%)	0

Note: Percentages may not add to 100% because of rounding to one decimal place.
a: Only 25 of the original 27 examiners answered this question. One did not continue from Phase 1 to Phase 2, and one, although participating in Phase 2, did not respond to this particular question.

Table 9: Ease of distinguishing between adjacent band levels before and after rating the samples

As shown in Table 9, the examiners tended towards the mid-point on both occasions. However, we can see that the examiners were less positive about distinguishing between band levels for the candidates in the samples than they were beforehand as only 16.0% reported finding it easy (rating of 1 or 2) after rating the samples, compared to 40.7% before rating the samples. Qualitative comments suggest that the language backgrounds selected for focus in this study may have been particularly challenging for some examiners:

E2: *Some accents that I am not used to hearing are more difficult to decipher than others. My ears are not as attuned to these sounds.*

E19: *It would be easier to differentiate if the accents were the same i.e. same linguistic background. My interview samples were largely from a group of candidates with accents with which I am not familiar.*

Other comments suggest that they may have also found it particularly difficult to distinguish between the band levels selected for this study (5, 6 and 7), as in:

E1: *The most difficult is for Band 7 and Band 5 where the descriptors cross some of one and some of another. I would rather have clear guidelines for each distinct area.*

E21: *Levels 6-7-8 are a little difficult sometimes.*

Further evidence that the examiners found it difficult to distinguish between the band levels selected for this study comes from two different data sources. Firstly, when asked in Questionnaire C (C2) whether they found any Pronunciation bands difficult to choose between when rating the candidates in the samples, all but two indicated that they found one or more distinctions problematic. Table 10. shows that distinctions between Bands 6-7 (6-7 and 6-7-8) and 5-6 were noted by the most examiners (54.1% and 37.5% of examiners respectively). The specific problems they cited when choosing between band levels will be discussed below with the findings addressing Research Question 2c).

Difficult band level decisions	Responses	
	n	% of examiners (n = 24)
4 – 5	5	20.8%
5 – 6	9	37.5%
5 – 7	2	8.3%
6 – 7	11	45.8%
6 – 7 – 8	2	8.3%
7 – 8	4	16.7%

Note: Although 24 examiners responded to this question, some indicated a number of band decisions that were difficult. The total of responses therefore does not add up to 24, and the percentages do not add up to 100%.

Table 10: Pronunciation bands examiners found difficult to choose between when awarding a Pronunciation score to the samples

The second indication that the examiners found the distinction between band levels 5, 6 and 7 problematic is the variation in Pronunciation scores they awarded to the samples. As outlined in the Methodology section, the 12 samples had already been awarded scores at band levels 5, 6 or 7 by IELTS Australia (four samples at each band level). Table 11 shows a breakdown of the number and percentage of scores awarded by the examiners. The figures in bold type are those where the score awarded by the examiners matched the IELTS-assigned score, and shaded cells indicate scores which differed by more than one band level. It can be seen that a range of scores from Band 3 to Band 8 were awarded, and that these were frequently different from the band level assigned by IELTS. At each band level, less than half of the scores awarded matched the IELTS-assigned score.

IELTS-assigned score	Scores awarded by examiners (n=26)												Total	
	3		4		5		6		7		8			
	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Band 5	1	1.0	29	27.9	**41**	**39.4**	28	26.9	5	4.8	0	0.0	104	100
Band 6	0	0	1	1.0	17	16.3	**44**	**42.3**	34	32.7	8	7.7	104	100
Band 7	0	0	4	3.8	23	22.1	42	40.4	**31**	**29.8**	4	3.8	104	100
Total	*1*	*0.3%*	*34*	*10.9%*	*81*	*26.0%*	*114*	*36.5%*	*70*	*22.4%*	*12*	*3.8%*	*312*	*100%*

Note: Percentages may not add up to totals or 100% because of rounding to one decimal place.

Table 11: Pronunciation scores awarded by examiners

Table 11 shows that the difficulty the examiners reported in distinguishing between band levels 6 and 7 (see Table 10) is reflected in the scores they awarded. Firstly, few examiners awarded a Pronunciation score of 7 to the Band 7 samples, suggesting a difficulty at this level. Less than one-third of the scores awarded to the Band 7 samples (29.8%) agreed with the IELTS-assigned score, and approximately one quarter of the scores differed by more than one band (differed by two bands: 22.1%; differed by three bands: 3.8%). Secondly, the examiners tended to award a score of 6 to the Band 7 samples, and actually awarded more scores of 6 than they did of 7 (40.4% of Band 7 samples were awarded a score of 6 while only 29.8% were awarded a score of 7). Conversely, they also awarded a score of 7 to a number of Band 6 samples.

Overall, Band 6 was the most commonly awarded score (36.5% or 114 of the total of 312 possible scores, compared to 26.0% and 22.4% for bands 5 and 7 respectively). Some qualitative comments made in response to various questions in Questionnaire C also suggest that there may be a tendency towards awarding a Band 6.

> E11 (C1): *Band 6 seemed the easiest and the most common. This is the band I usually give during the IELTS tests also.*
>
> E29 (C2): *To some extent I am hesitant to give a higher band score* [meaning above 6] *to candidates if there is still a noticeable accent even if they are actually quite easy to understand.*
>
> E11 (C5): *I would like* [the descriptors for] *bands 7 and 5 to be longer as often I find it difficult to differentiate. If I am confused, I often find myself choosing 6 as a default.*

The above findings were supported by the VP data. The Pronunciation scores awarded by each VP examiner are presented in Table 12. As this shows, there was also a tendency towards awarding a Band 6 score which accounted for half the scores awarded (12 of 24), and one examiner (VP5) actually awarded Band 6 to all of the samples she rated, and these included one Band 5, one Band 6 and two Band 7 samples.

Sample	Examiners					
	VP 1	VP 2	VP3	VP4	VP5	VP6
5A1	5		5			
5A2		5		6 (+1)		
5P1	6 (+1)				6 (+1)	
5P2		5				6 (+1)
6A1	8 (+2)				6	
6A2		6				7 (+1)
6P1			6	7 (+1)		
6P2			6	7 (+1)		
7A1			6 (-1)		6 (-1)	
7A2				6 (-1)		5 (-2)
7P1	8 (+1)				6 (-1)	
7P2		7				7

Note: (+) next to the score signifies that this score is higher than the IELTS-assigned score by the amount indicated
(-) signifies it is lower.

Table 12: Pronunciation scores awarded by VP examiners

The VP examiners seemed to have similar difficulties distinguishing between band levels:

- less than half of the 24 VP scores matched the IELTS-assigned scores (41.7%, n = 10)
- only two samples (5A1 and 7P2) were awarded the IELTS-assigned score by both examiners who rated them
- only one VP examiner (VP2) awarded IELTS-assigned band scores to all four samples that she rated.

Like Phase 2 examiners, VP examiners also seemed to have particular difficulties awarding a score of 7 to Band 7 samples, and tended instead to award a score of 6. Of the four samples where neither VP examiner score matched the IELTS-assigned score, three were Band 7 samples (7A1, 7A2, 7P1) and each was awarded at least one score of 6.

In summary, both the qualitative and quantitative data suggest that the examiners had some difficulty distinguishing between the band levels selected for this study. The scores they awarded ranged from Band 3 to Band 8, and frequently differed from the IELTS-assigned scores. The examiners reported that the distinction between Bands 6 and 7 was particularly problematic, and while examiners awarded a Band 6 to the samples most often, they were least likely to award a Band 7 score.

4.5 Research Question 2B)

Research Question 2b): When using the revised IELTS Pronunciation scale to award a pronunciation score to candidates at band levels 5, 6 and 7, which features of pronunciation do examiners take into consideration?

As discussed above (see Table 8), before rating the samples, Phase 2 examiners indicated that they considered intelligibility and listener effort to be the most important features when awarding a Pronunciation score. However, when actually rating the samples, it seems that other features might have also been important, particularly concrete features related to connected speech such as intonation, stress and rhythm. For example, comments made after rating the samples suggest that some examiners might have taken different concrete features into consideration when awarding scores at different levels, as in the following responses related to two different questions in Questionnaire C (C2 and C4). These suggest that when awarding a Band 7 or higher these examiners considered concrete features related to connected speech, such as rhythm, intonation and stress to distinguish these bands from those lower on the scale.

E21 (C2): *A high 6 is very close to a 7, with the impression of stress and intonation often making the difference.*

E23 (C4): *Individual sounds of course, and around Band 7 rhythm (esp. w. Indian/Pakistani speakers) is important.*

E22 (C2): *On the higher levels I have difficulty in distinguishing between 6, 7 or 8. For example 6- can generally be understood; whereas 8- can be easily understood – what lies between?? It must come down to intonation and stress.*

Comments from VP examiners in Phase 3 of the study provided insight into the features they took into consideration when actually rating the samples. In the rating stage of Phase 3, immediately after awarding a Pronunciation score, they summarised their reasons for choosing that particular score. Later, in the review stage, they provided verbal reports on the features that contributed to their assessment.

The frequency with which the VP examiners mentioned different features in the rating stage is presented in Table 13. From this we can see that the features most often mentioned were two concrete features related to connected speech, intonation (in 83.3%) and chunking (in 75.0%). This contrasts with the features that Phase 2 examiners identified before the rating task as most important, that is, intelligibility and listener effort (see Table 8). In fact, as shown in Table 13, the VP examiners mentioned intelligibility in only 54.2% of the summaries, which is less frequently than they mentioned a number of other features.

This difference may relate to the effect of the task, since VP examiners tended to rely on the descriptors listed at the relevant band levels and 'tick them off' one by one when giving their summaries, as in:

> VP2/5A2: *OK that speaker originally I was looking at a 4 because uhm he seemed to be quite hesitant but then I bumped him I had a look at number 6 because he seemed to fulfil all the first three parameters of Band 4. He had some acceptable phonological features, some pretty good chunking once he got warmed up but quite a few lapses in overall rhythm. His intonation and stress wasn't too bad. It was a little bit choppy sometimes. The reason I didn't give him a 6 was because he couldn't generally be understood throughout with much effort. He needed a bit of effort to be understood but not as much effort as he would have if he was a number 4. Individual words and phonemes may be pronounced but this causes only occasional lack of clarity* [reading from descriptors on scale]. *It was more frequent than an occasional. That's why I gave him a 5.*

Feature	VP examiner summaries in which the feature was mentioned (n = 24 [a])	
	n	%
Intonation	20	83.3%
Chunking	18	75.0%
Listener effort	17	70.8%
Stress[b]	16	66.7%
Rhythm	16	66.7%
Intelligibility	13	54.2%
Clarity	10	41.7%
Sounds	10	41.7%
Speech rate	6	25.0%
Accent	5	20.8%

[a] Number of VP summaries (6 VP examiners X 4 samples each).
[b] Includes word stress and sentence stress as comments did not always differentiate between the two.

Table 13: Features mentioned by VP examiners summarising their reasons for awarding a Pronunciation score

Since the VP examiners most often awarded scores at Band 5 and Band 6 (17 out of 24; see Table 12), when relying on the descriptors in this way, they would have made reference most frequently to features described in detail at Bands 4 and 6, and intelligibility is not referred to specifically in the descriptors at these levels. On the other hand, the Phase 2 examiners represented in Table 8 identified features from a list provided, and this list included intelligibility.

When providing verbal reports on the features that contributed to their assessment of the candidates' pronunciation during the review stage of Phase 3, however, intelligibility was among the top three features most frequently mentioned by the VP examiners. Table 14 provides a breakdown of the features mentioned, when they paused the recording to comment on anything that contributed to their assessment of the candidate's pronunciation (henceforth referred to as a review turn). From this it seems that, overall, phonemes were commented on the most, followed by intonation, intelligibility and so on. However, it is also clear that there was considerable variability among the six VP examiners.

4.5.1 Variation among examiners

As shown in Table 14, although most comments related to phonemes, the vast majority (64 of 76) of these came from just one examiner, VP3. Similar variation was found in what examiners commented on the most: while VP examiners 1 and 2 commented on intelligibility the most, VP5 and VP6 noted intonation most often and VP4 most commented on rhythm.

Feature		VP examiners' years of experience and number of review turns											
		VP1 (1.5yrs)		VP2 (0.2yrs)		VP3 (10 yrs)		VP4 (5yrs)		VP5 (2.5yrs)		VP6 (2yrs)	
		31		81		104		21		25		50	
	total	n	%	n	%	n	%	n	%	n	%	n	%
Phonemes	76	1	2.4	2	1.7	64	**45.4**	2	4.4	3	8.8	4	4.8
Intonation	66	2	4.8	15	12.4	8	5.7	5	11.1	11	**32.4**	25	**29.8**
Intelligibility	57	11	**23.8**	21	**17.4**	15	10.6	4	8.9	3	8.8	4	4.8
Word stress	53	3	7.1	12	9.9	20	14.2	3	6.7	2	5.9	13	15.5
Rhythm	51	5	11.9	13	10.7	11	7.8	10	**22.2**	0	0.0	12	14.3
Sentence stress	46	6	14.3	17	14.0	5	3.5	9	20.0	6	17.6	3	3.6
Listener effort	31	4	7.1	16	13.2	5	3.5	0	0.0	2	5.9	5	6.0
Chunking	30	2	4.8	9	7.4	3	2.1	3	6.7	1	2.9	12	14.3
Speech rate	25	1	2.4	10	8.3	4	2.8	7	15.6	2	5.9	1	1.2
Clarity	24	4	9.5	4	3.3	6	4.3	2	4.4	4	11.8	4	4.8
Accent	8	5	11.9	2	1.7	0	0.0	0	0.0	0	0.0	1	1.2
Total	467	42	100	121	100	141	100	45	100	34	100	84	100

Note: The figures in bold type in the shaded cells are the features mentioned at the most review turns by each VP examiner.

Table 14: Pronunciation features commented on in VP review turns by each examiner

As shown in Table 14, VP examiners varied considerably in the number of review turns they had (ranging from 21 by VP4 to 104 by VP3), and the features they noticed and commented on in those review turns suggesting variation in the features they felt warranted comment.

- As discussed above, with the exception of VP3, examiners commented very little on phonemes. Thus while these constituted 45.4% of VP3's comments, only 1.7% and 2.4% of comments by VP2 and VP1 respectively related specifically to phonemes.
- Although intonation was mentioned by VP5 and VP6 in 32.4% and 29.8% of their turns respectively, VP1 only commented on this feature in 4.8% of her turns.
- Although rhythm was the feature that VP4 commented on most frequently (in 22.2% of her turns), VP5 did not mention it at all.
- While VP5 mentioned word stress in 5.9% of her turns, VP6 mentioned it in 15.5% of hers.
- Similarly, the number of turns in which chunking was mentioned ranged from 2.9% (VP5) to 14.3% (VP6).

Some of this variation can be explained by the fact that each VP examiner listened to only four of the samples and a different mix in each case (see Table 12). However, considerable variability was found even when two VP examiners were rating the same sample, as shown in the following analysis.

As noted in the Methodology section, each sample was rated by two VP examiners. Of the 12 samples, only two were awarded the IELTS-assigned score by both VP examiners (5A1, examined by VP1 and VP3, and 7P2, examined by VP2 and VP6; see Table 12). These were therefore selected for close comparison. Table 15 shows the number of review turn comments made on each feature of pronunciation by each of the four examiners for these samples and illustrates considerable variation in the extent to which the two VP examiners for each sample noticed and commented on features, even though they were commenting on the same sample. VP3 made considerably more review turns than VP1 (32 compared to eight) on sample 5A1, and while VP3 focussed chiefly on the concrete features of phonemes (n=17, 38.6%), VP1 only mentioned them once and seemed to focus most on global judgements of intelligibility (n=5, 29.4%). VP3 also commented on some features not mentioned at all by VP1 (stress at word level, intonation and chunking), and VP1 mentioned rhythm once while VP3 did not mention it at all. However, both awarded the same score to 5A1, even though VP3 attended much more to concrete features while VP1 relied more on global judgements. VP3 was considerably more experienced than VP1, with 10 years as an IELTS examiner compared to 1.5 years. She also had experience as an examiner trainer, and so it is possible that she had greater expertise in identifying and articulating different features.

Table 15 also shows that there were considerable differences in the features identified by the two examiners who rated the sample 7P2. Here the difference in the number of review turns made by each VP examiner was not so extreme (19 compared to 13), but VP2 (the least experienced with only two months' experience) focussed most on global judgements of intelligibility (n=7, 21.9%) and listener effort (n=6, 18.8%), while VP6 (with two years' experience) only mentioned these features once and not at all respectively. In contrast, VP6 seemed to focus most on intonation (n=7, 25.9%), a feature that VP2 only mentioned once. Although both awarded the same score, they seem to have arrived at it by a slightly different route, related again perhaps to their experience and expertise in identifying and talking about different features.

	Sample 5A1				Sample 7P2			
Examiners (number of review turns)	**VP1** (n=8)		**VP3** (n=32)		**VP2** (n=19)		**VP6** (n=13)	
Features of pronunciation	n	% of review turns	n	% of review turns	n	% of review turns	n	% of review turns
Accent	0	0.0	0	0.0	2	6.3	1	3.7
Chunking	0	0.0	2	4.5	2	6.3	5	18.5
Clarity	2	11.8	4	9.1	1	3.1	2	7.4
Listener effort	2	11.8	3	6.8	6	18.8	0	0.0
Intelligibility	5	**29.4**	8	18.2	7	**21.9**	1	3,7
Intonation	0	0.0	2	4.5	1	3.1	7	**25.9**
Phonemes	1	5.9	17	**38.6**	1	3.1	2	7.4
Rhythm	1	5.9	0	0.0	2	6.3	3	11.1
Speech rate	0	0.0	0	0.0	2	6.3	1	3.7
Stress at word level	0	0.0	4	9.1	3	9.4	3	11.1
Stress at sentence level	3	17.6	3	9.4	3	9.4	1	3.7
Total	**17**	**100.0**	**44**	**100.0**	**32**	**100.0**	**27**	**100.0**

Note: The figures in bold type are the feature mentioned at the most review turns for each VP examiner.

Table 15: Frequency of comments on features of pronunciation at review turns when same score awarded by both VP examiners

It is possible that on some occasions the VP examiners in Table 15 were in fact attending to the same features in the speech signal but describing them in different ways according to their level of technical expertise. For example, some of VP1's comments about intelligibility may have been related to the phoneme errors commented on by VP3, and some of VP2's comments about listener effort may have been related to word stress issues commented on by VP6. However, although this did happen on some occasions (see Category 2 below), it was not always the case. As discussed in the following analysis, even when the VP examiners paused the recording of the same sample to comment on the same section of speech, they did not always comment on the same features in the speech signal. In addition, even when commenting on the same feature, they did not always describe it the same way and sometimes disagreed whether the candidate was using features correctly or not.

Not only did the review turns made by each VP examiner vary in number but they also often related to the different stretches of speech. There were only 16 occasions (eight for each sample) when both VP examiners paused the recording in the same place to comment, that is, when their review turns corresponded. Analysis of their comments made at these turns revealed that there were only five out of the 16 cases in which examiners were in total agreement (see Category 1 below).

Corresponding review turns fell into four different categories.

Category 1: They commented on the same feature/s and their description was similar (n=5).

For example, when commenting on the same section of speech in sample 5A1 where the candidate used sentence stress appropriately, both VP examiners commented on this, as in:

VP1: *His stress is better here* [referring to the previous comment indicating that the candidate had not placed stress properly on the right words for overall meaning].

VP3: *'But when I WORK' so he's got the stress.*

Category 2: They commented on the same feature/s but their description was different (n=4).

For example, when commenting on mispronounced phonemes in a section of speech in sample 5A1, VP1 used more general terms than those used by VP3. It should be noted, that although it is likely that both examiners were commenting on the word stress and phoneme issues in this section of speech, VP3 actually paused the recorder eight times to comment on specific features while VP1 paused it once at the end of the section and commented in general, as in:

VP1: *Lots of misunderstanding here. Some effort needed to understand catches of this little sentences and phrases what he's saying here.*

VP3: 1. *'I will retain my country' not 'return my country'* [referring to vowel production in *return*].

2. *'The technology /dɛfləbəm/' Can't tell what that was.*

3. *'A lot of beople' not 'people' So again it's that sort of /b /p/ /t/ all of those kind of sounds.*

4. 'Lost the /ʒɒbz/'

5. *'/fjuʃə/' not 'future'*

6. *'We have reboard?'* [Had trouble identifying the word *robot* and copied the candidate's incorrect stress pattern.]

7. *Now so that's a patch that would bring it down from a 6 to a 5 so that's the bits of the 4 coming in there that bring it down* [indicating that it's hard to understand].

8. *I think he said 'human body' there or 'human beddy'.*

This example suggests differences between the two examiners in the detail and precision with which they were able to comment. As mentioned earlier, this could be related to their level of experience as IELTS examiners.

Category 3: They commented on the same feature but disagreed as to whether it was used appropriately (n=2).

This occurred on one occasion for each sample. In sample 7P2, the two examiners did not agree on the quality of the chunking in a section of speech where our analysis indicated that it was not used appropriately:

VP2: *His chunking and his rhythm was [sic] a little bit screwed up there I think while he was trying to find what he was going to say next.*

VP6: *Slight problem with intonation here but generally speaking again, chunking is good.*

Category 4: They commented on different features (n=5).

For example, when commenting on the same section of speech in sample 7P2 where our analysis showed misplaced stress on the word *photography*, VP 2 commented that sentence stress was appropriate and VP6 commented that there was a problem with the word stress, ie they stopped the recording in the same place but commented on different aspects of stress:

VP2: *He's got the stress there 'and they work SO hard'.*

VP6: *Again 'photography'* [referring to a previous comment regarding a word stress issue in the word '*photographer'*].

In summary, the features of pronunciation the VP examiners mentioned most when summarising their reasons for awarding a particular Pronunciation score were two concrete features related to connected speech – intonation and chunking – followed by the global feature, listener effort. However, when providing verbal reports on the features that contributed to their perception of the candidates' pronunciation, there was variability in both the number of review turns and in the features they commented on at those review turns. This suggests that the VP examiners were noticing and commenting on different features during this process. Furthermore, even when review turns did correspond, that is, where both VP examiners paused the recording to comment on the same section of speech, the features they commented on were not necessarily the same, and even when commenting on the same feature, they did not always describe it the same way and sometimes disagreed whether the candidate was using it correctly or not. It was also evident, as discussed below, that the examiners' use of terms referring to global features of pronunciation was not always consistent.

4.5.2 Global features of pronunciation: clarity, intelligibility and listener effort

It is important to note here that, although we have followed previous studies in counting comments on different features (see for example, Hubbard et al 2006), it was by no means always clear exactly which feature VP examiners were commenting on or to distinguish clearly between judgements on the more global aspects of clarity, intelligibility and listener effort. In comments where VP examiners said they could or could not understand a particular section of speech, or noted how hard or easy something was to understand, there was a clear basis for identifying intelligibility (ie how much they understand) and listener effort (ie how hard or easy something is to understand) respectively as the focus of interest. However, although previous studies have found these to be independent dimensions (Derwing & Munro 1997; Munro & Derwing 1995a), they are very closely related and comments often referred to both. For example:

VP6/7A2: *Now I would say this needs a little bit of effort to understand cos the first time I listened I actually didn't understand him.*

VP6/7A2: *This is where I cannot understand uhm the candidate so I would say that I need some effort to understand him.*

Similarly, when the VP examiners commented on clarity it was not always clear what was covered. In their summaries, the term mostly (in seven out of 10 mentions) related to the mispronunciation of features at the word level. Sometimes they referred to concrete features such as phonemes, as in:

> VP3/5A1: *so things like 'celery' for 'salary', 'high' for 'hired', 'bu- sometime' so obviously missing off quite a lot of endings of words which causes more than uhm more than the occasional lack of clarity*.

Sometimes references remained general terms, as in:

> VP1/5A1: *some mispronounced words uhm and causing the occasional lack of clarity.*

This meant that VP examiners did not always unpack the specific concrete features at the root of the perceived lack of clarity.

There were also instances where the term seemed to include a reference to global judgements of listener effort and/ or intelligibility, as in:

> VP3/5A1: *'/deleɪtəd/' '/deleɪtəd/' That sort of combination 'that are related' uhm he squishes it together which makes it difficult to understand or makes it slightly unclear.*
>
> VP6/7A2: *Now this is one particular instance that could be considered as difficult to understand or occasional lack of clarity.*

This is not necessarily surprising, as mispronunciations can lead to difficulties in interpretation of meaning. However, it was not always clear exactly what concrete features VP examiners were referring to when they used the term clarity. At times, examiners also commented on features not directly related to the Pronunciation scale, as discussed below.

4.5.3 Consideration of features not included in the revised Pronunciation scale

Although VP examiners were only asked to comment on features related to the revised Pronunciation scale, they also made comment on a range of other features, suggesting that these might also have played a role in their judgements. At 46 review turns (14.7% of the total 312), VP examiners included comment on features and/or used terms not included in the descriptors or key indicators for the revised Pronunciation scale (henceforward non-P scale comments). Furthermore, at 23 of these (7.4% of the total 312) the non-P scale comment was the only comment made.

Non-P scale comments fell into three main categories.

1. Comments related to a different scale in the Speaking Test (n=18). For example, on Grammatical Range and Accuracy:

   > VP5/7A1: *She leaves out a few uhm definite articles and things like that but it still doesn't sort of really trap her - doesn't seem to trip her up that much.*

   And Fluency and Coherence (see also discussion in Section 4.6.2):

   > VP1/5A1: *Yeah his rate of speech, that's sort of bringing him down from a 6 down to a 5 because of the the – the speech is quite slow and hesitant in a way.*

2. Comments that involved non-IELTS terms (n=18). In some of these (n=7), the relationship to pronunciation could be inferred as in:

   > VP2/5P2: *He's mumbling a little bit there.* (Clarity)

VP4/6P2: *It feels like she needs to take a breath.* (Chunking)

In most (n=11), however, the nature of the pronunciation issue being commented on was unclear, as in:

VP5/7A1: *It may not come out - roll off her tongue like that to start with.*

VP5/7P1: *He seems to have you know his tone is a little bit sharper now.*

3. Affective comments that reflected on what the candidate might be feeling or thinking, (n=15). For example:

VP1/5P1: *I mean he seems to be sort of happy with what he's saying. He seems to know what he wants to say but – so he seems to come across a bit more confident with what he's saying.*

VP2/7P2: *Now that he's warmed up a little bit I'm going to be moving up to probably about a 6 and be focussing around that area.*

(Note: at some review turns there was more than one non-P scale comment, so the sum of the totals for each category is more than the number of review turns where non-P scale comments were made.)

The frequency of such non-P scale comments suggests that some examiners may have blurred the boundaries between Pronunciation scale and non-P scale features, suggesting that these latter may also have played a role in their assessments of pronunciation. The extent to which examiners made only non-P scale comments at review turns varied enormously, ranging from none (VP4 and VP6) to 13 (VP5). This means that for VP5, non-P scale review turns accounted for over half of her total number of review turns (52.0%; 13 out or 25) and that she made more non-P scale comments than comments on intonation, the pronunciation feature she mentioned the most (13 compared to 11, see Table 14). Interestingly, as noted earlier (Table 12), this examiner was limited in the range of scores she awarded: she only awarded Band 6 Pronunciation scores, and this matched the IELTS-assigned score for only one sample.

As indicated above, non-P scale comments related to the other scales in the Speaking Test were made at 18 review turns. The most common of these (n = 12) were comments related to the Fluency and Coherence scale, and as discussed below, the overlap between Pronunciation and Fluency and Coherence descriptors was one of the main problems the examiners reported regarding the use of the revised Pronunciation scale.

4.6 Research Question 2C)

Research Question 2c): When using the revised IELTS Pronunciation scale to award a pronunciation score to candidates at band levels 5, 6 and 7, what problems do examiners report regarding use of the scale?

Before rating the samples in this study, the responses of Phase 2 examiners to Questionnaire A indicated that they were largely positive about how easy the descriptors and increased number of band levels were to use (see Research Question 1). However, analysis of the responses they gave to Questionnaire C after rating the samples and of the comments made by VP examiners in Phase 3 of the study revealed two major areas of concern: (a) the descriptors at Bands 3, 5 and 7, and (b) the overlap between the Pronunciation scale and the Fluency and Coherence scale.

4.6.1 The descriptors at Bands 3, 5 and 7

In response to a number of different questions in Questionnaire C, the majority of Phase 2 examiners (19 out of 26, 73.1%) indicated at some stage that they would like more specific details in the Pronunciation descriptors at Bands 3, 5 and 7 or felt that the current wording at these levels was difficult to interpret. In particular, issues were raised concerning the interpretation of 'positive features' and 'some but not all'. In response to C2, which elicited any bands they found difficult to choose between, seven examiners referred specifically to the difficulty of using Bands 5 and 7. For example:

> E14: *5 and 4 is always hard. eg '5' refers to all the positive features of Band 4, yet there aren't many!!*
>
> E11: *5-6 especially, as descriptions for Band 5 are quite minimal, similar for Band 7. 6-7.*
>
> E13: *6 and 7: It is not clear what qualifies as a '7'.*
>
> E18: *6- 7-8 because there is no specific Band 7 indicator, it makes it a little more difficult to assess.*

In C3, examiners were asked to choose between the following three statements regarding the rationale for awarding a band score of 5 and they were then invited to comment on their answer.

1. The candidate displays all the features of 4 and most of the positive features of 6.
2. The candidate displays all of the features of 4 and all but one of the positive features of 6.
3. The candidate appears to be mid-way between a 4 and a 6.

It was evident from their responses to Q3 that there was some confusion about the interpretation of the wording 'some but not all' in the new 'in between' band levels. Of the 26 examiners who responded to this question, only two selected Statement 2, the interpretation that best fits the guidelines given in the instructions to examiners (IELTS 2008a). Their comments suggest that rather than identifying all the features of 4 and all but one of the positive features of 6, in practice they focus on a specific feature when determining the score:

> E5: *I tend to use the 'chunking' descriptor more as a benchmark.*
>
> E29: *I tend to focus more on 'can be generally understood...' descriptor even though the other descriptors are also considered.*

Ten examiners selected Statement 1 indicating they thought a speaker should be awarded a Band 5 if a candidate displays 'most of the positive features of 6'. For example:

E6: *I would award a 5 if the candidate achieves 2 or more (but not all) of the Band 6 criteria.*

E19: *Some, but not all could fall under any of the above definitions. This is poorly expressed and too arbitrary especially in something as important as pronunciation.*

And 10 examiners selected Statement 3: 'the candidate appears to be mid-way between a 4 and a 6'. For example:

E16: *Which are the positive and which are the negative? It would be good to have a 'star' indicating the positive features. e.g. 4) attempts to use intonation but control is limited. Is this positive or negative?*

E22: 1+*2 above do not suggest a mid-way mark between 4+6.*

E26: *The candidate should display all positives of 4 and at least one positive of 6 (not 'most' or 'all but one').*

The remaining four examiners either chose two alternatives or amended one of the options to fit their view of what the correct answer should be.

In their responses to C5, which asked them to comment on the length of the Pronunciation descriptors, just over half of the examiners (14 or 53.9%) felt that the descriptors were the right length, six (23.1%) indicated they should be longer, and six explicitly commented that the descriptors for Bands 5 and 7 were inadequate. The length and wording of the descriptors for Bands 3, 5 and 7 were the subject of negative comments by nine examiners, and a further two suggested more 'options' or 'guidance' could be given for these levels. Some were quite critical in their evaluation of these new band levels and two referred to these bands as 'cop outs'. For example:

E19: *The new descriptor bands, especially 5 and 7 are inadequate and a bit of a 'cop out', especially given that many, if not most candidates, will fall in this range. We waited so long for these new bands, and were so accustomed to the usefulness of the band levels for the other descriptors, that the new band came as a great disappointment.*

Others indicated a desire for more specific descriptions at the new band levels and reported having issues with the concept of 'positive features'. For example:

E14: *I don't like the 'displays all positive features of Band X and some but not all positive features of Band Y.' Too confusing in time pressure situation.*

E16: *In the new Bands 9-7-5 I would like the positive features to be listed of the bands above and below eg for a '7' I would like the positive features of the 8 and 6 to be written as the descriptor.*

When asked to comment on what they didn't like about the revised Pronunciation scale (C11), 15 of the 21 examiners who commented, referred to the need for greater specificity, and most of these referred to more detail at Bands 3, 5 and 7. For example:

E2: *As mentioned, 3, 5 and 7 could be expanded more.*

E16: *I'd like descriptors listed in the new Bands.*

E19: *This is insufficient detail in Bands 3, 5 and 7.*

E21: *Levels 5 and 7 are defined by reference to other levels which is not always easy.*

E27: *No detail for Bands 3, 5 and 7.*

Further evidence of some concern over the wording of these descriptors came from responses to C12, which asked for comments on how the revised scale could be improved. Here 14 of the 24 examiners who responded suggested altering the wording of the descriptors for Bands 3, 5 and 7. For example:

E10: *It can be changed to be 'mid-way' or more descriptors given.*

E13: *Again, being more explicit in the in-between Bands - 3, 5, 7.*

E14: *Get rid of the 'displays features of X, but not all features of Y.' Replace with more easily assessable descriptors.*

E18: *A bit more detail for Bands 3, 5 and 7.*

E20: *The descriptors for odd numbers should be more explicit. 'some, but not all,...' needs to be clearer.*

E26: *Maybe, put negative feature in bold for each band to distinguish them.*

It seems, therefore, that although the examiners preferred the revised scale to the old scale, they still had some concerns and confusion over the interpretation of the descriptors in the added bands, and many felt the need for greater specificity. There was also some concern expressed about the overlap between the Pronunciation scale and the Fluency and Coherence scale.

4.6.2 The overlap between the Pronunciation scale and the Fluency and Coherence scale

Overlap between the Pronunciation scale and Fluency and Coherence scale was a consistent theme both in comments made by Phase 2 examiners and at review turns by the VP examiners in Phase 3. Eight of the 26 Phase 2 examiners commented at some point in Questionnaire C on difficulties related to differentiating between these two scales and managing the perceived overlap. For example:

E2 (C7): *I wonder if speech rate should be under Fluency and Coherence rather than under Pronunciation.*

E10 (C7): *I know repetition is covered in F& C, but I find it affects intelligibility in some candidates.*

E13: (C1): *I found a lot of speech samples had problems with rhythm. In terms of IELTS scoring, this is closely tied to FC - how quickly a candidate speaks and their ability to chunk language.*

E14 (C13): *I still don't always know how to assess, accurately, if I've given a candidate a 6 for coherence, but I really feel they only deserve a 4 for pron. Should I review coherence in light of v. poor pron? I sometimes struggle with this.*

Further insight into concerns about the overlap between these two scales emerged from review turn comments by the VP examiners. As noted above (see Research Question 2b), at 12 review turns, VP examiners commented on features of Fluency and Coherence rather than Pronunciation, as in:

VP1/5A1: *Yeah his rate of speech, that's sort of bringing him down from a 6 down to a 5 because of the the- the speech is quite slow and hesitant in a way.*

Further comments made by two of the VP examiners at the end of their verbal protocol sessions also indicated that they found the relationship between the two scales very close.

VP2: *It's kind of hard to separate the fluency from the pronunciation. I mean you have to really conscious to try to separate them, I mean they're not completely stand alone but it's kind of hard just to mark the pronunciation without a bit of fluency bias… fluency is quite closely related to pronunciation cos if someone speaks exceptionally slowly but they pronounce things really well you tend to mark them down because the speed isn't quite up to it.*

VP3: W*hen you're looking at a profile if you get someone who, like for example, 8, 6, 6, 4, across the board, you'd look at it and you'd go, 'that's nigh on impossible' because to have good fluency and coherence, it's the stresses, the pauses, the hesitations and all of that, and that of course has an effect on the syllable timing and the stress timing and the rhythm which influence the pronunciation so if you looked at that jagged profile you'd listen to it again because you'd go, 'that's just weird'.*

4.7 Summary of findings

In general, the Phase 2 examiners preferred the revised Pronunciation scale to the previous one, and were largely positive about how easy it was to use the descriptors and increased number of band levels. They reported feeling confident about assessing the different features of pronunciation covered in the Pronunciation scale descriptors, and most confident about making global judgements of intelligibility and listener effort, which were the features they considered to be the most important when awarding a Pronunciation score. However, when actually rating the samples in this study, they had some difficulty distinguishing between the different band levels, and awarded Pronunciation scores ranging from Band 3 to Band 8 to candidates with IELTS-assigned scores of 5, 6 or 7. They reported that the distinction between Bands 6 and 7 was particularly problematic, and seemed reluctant to award a Pronunciation score of 7 to the Band 7 samples. Band 6 was the most commonly awarded score, and this was related in part to the tendency to award a Band 6 rather than 7 to the Band 7 samples.

The difficulty distinguishing between Bands 5, 6 and 7 when rating the samples was also reflected in VP data, where Band 6 was the most commonly awarded score and less than half the scores awarded matched the IELTS-assigned scores. VP examiners rating the samples reported two concrete features related to connected speech – intonation and chunking – to be the most important, followed by the global judgement of listener effort. However, when providing verbal reports on the features that contributed to their assessment of the candidates' pronunciation, they varied in the features they noticed and commented on. This variation was evident even when examiners commented on the same section of speech: they did not necessarily mention the same features, and even when commenting on the same feature, they did not always describe it in the same way, and sometimes disagreed as to whether the candidate was using it correctly or not. It was also evident that their use of terms referring to global features of pronunciation was not always consistent, and that some VP examiners may have been influenced by features not included in Pronunciation scale descriptors when awarding Pronunciation scores.

Two areas of concern about the revised Pronunciation scale were identified: (a) the specificity of the descriptors at Bands 3, 5 and 7, and (b) its overlap with the Fluency and Coherence scale.

5 DISCUSSION

5.1 Examiner attitudes to, and use of, the scales

The aim of this study was to explore how examiners view the revised Pronunciation scale in general, and to investigate their use of the scale to award scores to speakers from two different language backgrounds at the crucial Pronunciation band levels of 5, 6 and 7. The findings suggest that, in general, these examiners preferred the revised Pronunciation scale to the previous version, felt confident about assessing the features covered in the descriptors and were largely positive about how easy the increased number of band levels were to use. Their general approval of the length and content of the scale suggests that it has avoided one of the operational dangers of scales of this kind – long and complicated descriptors that examiners find inaccessible. As Orr (2002) notes, the more complicated and detailed the descriptors in a scale, the less likely it is to be used consistently.

Although they viewed the revised Pronunciation scale quite positively, the examiners did have some difficulty distinguishing between the different band levels when using it to award scores to the samples in this study. The distinction between Bands 6 and 7 seemed to be particularly problematic, and they tended to award a score of 6 rather than 7 to the Band 7 samples. In addition, there was an overall tendency to gravitate towards awarding a score of 6, even though Band 5, 6 and 7 candidates were equally represented in the samples. Brown (2006, p 59) observed that examiners using the previous four-point scale tended to use Band 6 as the 'default' level because they were reluctant to award Bands 4 or 8. This tendency could have important real-world consequences for candidates taking the test for a range of gate-keeping purposes. It seems from the current findings, however, that the inclusion of the in-between bands in the revised Pronunciation scale has not necessarily laid this 'default' to rest.

A possible source of confusion which may have encouraged these tendencies was highlighted by the issues examiners had in interpreting the new in-between bands. As discussed in the Results section, there seemed to be confusion around how to interpret the wording 'some but not all' in the descriptors at Bands 5 and 7. This variation in interpretation may be explained in part by an apparent discrepancy in the IELTS documentation. The descriptors themselves state that 'some, but not all' of the positive features of band 6 must be present for a 5 to be awarded. However, the self-access training materials for examiners (IELTS 2008a) offer a slightly different definition. Some clarification of the exact intention or greater specification of what is intended at these band levels might therefore be helpful.

Another issue for examiners in using the Pronunciation scale appeared to be a perceived overlap with Fluency and Coherence. These two aspects of spoken English are closely related and it is difficult to separate out factors that combine to play a role in speaking proficiency. For instance, pausing appropriately so that words are grouped into meaningful 'chunks' is considered a feature of pronunciation (see for example Cauldwell 2003). Yet the number of pauses and the number of words between pauses have been used in research as a temporal measure of fluency (see Segalowitz 2010). Similarly, while some authors consider speech rate to be an aspect of pronunciation (see, for example, Iwashita et al 2008), it is included at some point in both scales in the IELTS descriptors. Some of the difficulty in separating these two scales, however, could also be related to a certain amount of overlap in documentation provided by IELTS on the different scales in the Speaking Test (IELTS 2008b), particularly as it relates to speech rate, hesitation and chunking. Although these are closely related and can be seen as production variables relating either to pronunciation or to a fluent and coherent performance, there is some repetition in the wording used in both scales, and this might add to the examiners' difficulties in separating out the two scales. These areas of perceived overlap seem to complicate the process of awarding a discrete score for pronunciation for some examiners.

5.2 Variation between examiners

Examiners varied in a number of respects, and some insight into this variation was given by the VP data: examiners varied not only in the score that they gave to the same sample, but also in the features of pronunciation they attended to in their rating and review stage comments. Differences among examiners were evident in the number and sections of the samples in which they noted features for comment, the features they chose to comment on, and how they described those features.

The assessment of speaking skills is notoriously challenging and a certain amount of variability is an inevitable part of the process (McNamara 1996, p 127). In his study of rating behaviour in the spoken test of the FCE, Orr (2002) also found considerable variability among examiners who sometimes rated a sample in a similar way but assigned different scores or, alternatively, awarded the same score but drew on different aspects of the scale and commented on different aspects of the speaker's performance. Commenting on the variation in what examiners attend to in his study, Orr (2002) concluded that 'for each rater there appears to have been a unique interaction of factors which led to the awarding of a score' (Orr 2002, p 151). While the context of the VP itself may help to explain some of the variation in what VP examiners chose to comment on (see, for example, Hubbard et al 2006), and the use of a scale with specific descriptors seems to have addressed this variation to some extent, it nevertheless seems that individual factors of personal interpretation, interest or expertise remain an issue.

A range of factors seems to have contributed to this variability. These included factors related to individual professional experience, expertise and preference, the nature of both spoken assessment in general and the nature of making assessments according to a scale in particular. It is worth noting that while examiners in this study completed the questionnaire and rating tasks on familiar territory and were allowed to review the sample recording as they might in a genuine test condition, they did not have the benefit of a face-to face encounter and only had access to one part of the spoken interview test on which to base their rating judgements and scores. This situation does not exactly mirror the test situation and it has been argued that examiners rate audio samples more severely (Taylor & Jones 2001, p 2). However, since the examiner scores varied in both directions at Bands 5 and 6, other factors are obviously important here.

Although the participants in the study were all trained and current IELTS examiners from a single centre, their teaching experience varied from three to 30 years and their experience as examiners from less than a year to 13 years. Although the research questions did not directly address the issue of the relationship between such factors and a tendency to score in a particular way, there was some indication from the VP data that examiner background may be important. As discussed in the Results section, it was not necessarily those who had the most experience as an examiner or as a teacher of English who awarded scores to the samples that most closely matched those assigned by IELTS. However, expertise seemed to play a role, at least in the precision with which certain concrete phonological features could be identified and described.

In the identification of phonemes, for example, one VP examiner was clearly knowledgeable in this area and had the expertise to identify and describe the issues, while others rarely commented on this feature in detail. How far they were in a position to comment explicitly on this feature was not entirely clear. As discussed earlier, this is an area in which many teachers lack confidence and even experienced listeners have difficulty in making judgements (Schmid & Yeni-Komshian 1999; Derwing & Rossiter 2003; Levis 2006; Macdonald 2002). When an examiner did not comment, we do not know whether this was because she did not notice a particular feature, whether she noticed it but felt she was not able to comment with sufficient expertise, or noticed it and felt that it was not worthy of comment.

5.3 The rating process and what examiners take into consideration

Although it is difficult to make generalisations about the rating process from verbal protocols as these inevitably represent some sort of an intrusion into the normal process (Brown 2007), there were indications that examiners tended to use the Pronunciation scale and key indicators as a basis for descriptions of pronunciation features in speech samples and as a checklist against which these are considered for awarding scores at different levels. As such, these offer them a discourse that they can use to articulate what they have noticed as they refer to descriptors listed at the relevant band levels and 'tick off' the features one by one, in a similar way to the process described in Brown (2007). This close use of the scale seems to have encouraged examiners to pay attention to a range of features when awarding a score and given them the framework within which to talk about the same aspects of a performance. In the questionnaire data, there was some convergence on which features they felt were important in assessing pronunciation. In the verbal protocol, too, examiners oriented to similar features noted in the scale, albeit to different degrees and with different emphases. To this extent, the scale seemed to have provided them with a script that they could follow when talking about or reflecting on a candidate's performance. This has been noted by Lumley (2005, p 311), who describes an assessment scale as offering 'language and modus operandi for raters to follow in describing their justifications'.

It is interesting that, despite the overt move in the descriptors to stress the importance of concrete phonological features and downplay the importance of global judgements, the Phase 1 examiners still rated global judgements related to intelligibility and listener effort as very important in making scoring decisions, although, in line with instructions accompanying the Pronunciation scale, they disregarded accent *per se* as an issue. One explanation for the popularity of global judgements of this kind might be that the majority of the examiners (21 out of 27) had had experience using the previous scale, and this may have still exerted some influence over the way they thought about (or at least reported thinking about) pronunciation. Another explanation might be the general and undemanding nature of such judgements in terms of technical expertise, that is, it is much easier to say that a stretch of speech is unintelligible or difficult to understand than it is to give a precise technical analysis of specific concrete problems. This is a strength, in that it allows an examiner to make an assessment even if they have little training in phonology, but also a weakness, in that it is allows considerable latitude in how such judgements are made. It is quite a complex matter to determine degrees of intelligibility or exactly what is meant by 'easily understood', and the VP data have illustrated that examiners may not always mean the same thing when they use terms such as 'unclear' or 'difficult to understand'.

Features such as clarity, intelligibility and listener effort provided a terminology which examiners used, but it was not always evident from the VP data that they saw these features in the same way or used them to refer to the same phenomena in the speech samples. In other words, these concepts seemed to allow a degree of license in what they covered, and this license could allow examiners to remain imprecise about exactly what they identified in the speech sample. Moreover, as Brown (2007) found, there can be differences among examiners in their level of tolerance for more global judgements such as those for comprehensibility. Closer definition of these global aspects of pronunciation might be helpful here. For example, although the concept of clarity occurs regularly in the descriptors in relation to word level features such as word stress and phonemes, it is not defined in the glossary. At times it was evidently used to refer to the precision of articulation of sounds in words but elsewhere it was used in a way that seemed closer to the concept of intelligibility, rather than articulatory accuracy. Of course, the two are related: unclear articulation can certainly make stretches of speech difficult to understand. It seemed, however, that the relationship between the two was not entirely clear for some examiners.

Thus the scale appears to offer a useful checklist for the assessment of pronunciation and may help to focus examiners' attention. As Lumley (2005 p 305) notes, since raters have only limited time to talk about their assessments during a VP, they are likely to make explicit reference to the scale to justify their scoring decisions 'because that is what they are required to do'. However, the scale also potentially offers examiners with a means of talking about samples in a way that appears focussed but which may mask a certain amount of variation and imprecision. Orr (2002), for example, reported that a third of the examiners in his study oriented towards global impressions and noted as a consequence 'the limitations of the rating scale and the training for focussing raters' attention on the components of communicative language ability and not its overall effect' (p 151).

A further issue found by Orr (2002) in his study of the processes of rating spoken performance was the frequency with which examiners commented on factors outside the scale. He concludes that the raters in his study did not understand 'the model of communicative language ability on which the rating scales are based' (p 152). While the insights from the VP phase of this study suggest that some examiners did stray outside the descriptors of the Pronunciation scale and occasionally used vague descriptions, this kind of off-topic comment seems to have been less of an issue. Iwashita et al (2008), also found this variability to be less of a problem, and rather, that the raters in their study weighed up several factors within the scales to reach a score. Individual variation notwithstanding, the use of the revised Pronunciation scale may have assisted the VP examiners in this study to stay on-topic.

6 CONCLUSION AND IMPLICATIONS

Fulcher, Davidson and Kemp (2011) argue that the use and interpretation of a scale depends on socialisation, that is, on how well examiners can be trained and encouraged to use and understand it in the way intended by the test developers and consistent with other examiners worldwide. While the findings of the study suggest that examiners are generally positive about the revised Pronunciation scale and use it as a focus for the examining process and the awarding of scores, they do not always seem to be clear about the descriptors at certain band levels and may benefit from professional development on how the certain features relate to each other and to spoken performances at different levels. We therefore make the following suggestions.

- Some revision be made to the descriptors at Bands 3, 5 and 7 so that specific, concrete features of performances are identified at these levels, or further guidelines be adopted which clarify how the current descriptors are to be interpreted.
- Instructions in training documentation be clarified to ensure consistent interpretation of the Band descriptors 3, 5 and 7.
- Guidelines be developed to assist examiners to distinguish between similar features in the Pronunciation scale and Fluency and Coherence scale.
- Ongoing professional development, re-certification and moderation of examiners target issues in pronunciation and the rating process, specifically:
  - the nature of the scale and how to recognise the features of pronunciation
  - the standardisation of scores and how they reflect the presence or absence of particular features
  - the relationship between the Pronunciation and Fluency and Coherence scales.
- Examiner selection processes ensure a minimal level of expertise in pronunciation.

REFERENCES

Bent, T, Bradlow, A and Smith, B, 2007, 'Segmental errors in different word positions and their effects on intelligibility of non-native speech' in *Language Experience in Second Language Speech Learning*, eds O-S Bohn and MJ Munro, John Benjamins Publishing Company, Amsterdam, pp 331-347

Birrell, R and Healey, E, 2008, 'How are skilled migrants doing?', *People and Place*, vol 16. no 1, pp 1-19

Boyd, S, 2003, 'Foreign-born teachers in the multilingual classroom in Sweden: the role of attitudes to foreign accent', *International Journal of Bilingual Education and Bilingualism*, vol 6, no 3/4, pp 283-295

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 41-65

Brown, A, 2007, 'An investigation of the rating process in the IELTS oral interview' in *IELTS collected papers: research in speaking and writing assessment*, eds L Taylor and P Flavey, Cambridge University Press, Cambridge, pp 98-139

Brown, A and Taylor, L, 2006, 'A worldwide survey of examiners' views and experience of the revised IELTS Speaking Test', *Research Notes*, vol 26, pp 14-18

Cauldwell, R, 2002, *Streaming speech: listening and pronunciation for advanced learners of English*, speechinaction, Birmingham, UK

Derwing, TM and Munro, MJ, 1997, 'Accent, intelligibility, and comprehensibility: evidence from four L1s', *Studies in Second Language Acquisition*, vol 19, no 1, pp 1-16

Derwing, TM and Munro, MJ, 2005, 'Second language accent and pronunciation teaching: a research-based approach', *TESOL Quarterly*, vol 39, pp 379-398

Derwing, TM and Rossiter, M, 2003, 'The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech', *Applied Language Learning*, vol 13, no 1, pp 1-17

De Velle, S, 2008, 'The revised IELTS Pronunciation scale', *Research Notes*, vol 34, pp 36-38

Fayer, JM and Krasinski, E, 1987, 'Native and nonnative judgments of intelligibility and irritation', *Language Learning*, vol 37, no 3, pp 313-326

Field, J, 2005, 'Intelligibility and the listener: the role of lexical stress', *TESOL Quarterly*, vol 39, pp 399-424.

Fulcher, G, Davidson, F and Kemp, J, (2011), 'Effective rating scale development for speaking tests: Performance decision trees' in *Language Testing*

Hahn, LD, 2004, 'Primary stress and intelligibility: research to motivate the teaching of suprasegmentals', *TESOL Quarterly*, vol 38, no 2, pp 201-223

Hansen Edwards, JG and Zampini, M, (eds), 2008, *Phonology and second language acquisition*, John Benjamins Publishing Company, Amsterdam/Philadelphia

Hubbard, C, Gilbert, S and Pidcock, J, 2006, 'Assessment processes in speaking tests: a pilot verbal protocol study', *Research Notes*, vol 24, pp 14-19

IELTS, 2008a, *IELTS Speaking Test. Self-Access Re-training Set for the Revised Pronunciation Scale*, IELTS, Cambridge

IELTS, 2008b, *IELTS Speaking Test. Instructions to IELTS examiners*, IELTS, Cambridge

IELTS, 2010, *Examiner information,* accessed 21 October 2010, from <www.ielts.org>

Iwashita, N, Brown, A, McNamara, T and O'Hagan, S, 2008, 'Assessed levels of second language speaking proficiency: how distinct?', *Applied Linguistics*, vol 29, no 1, pp 24-49

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 245-269.

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

MacDonald, S, 2002, 'Pronunciation: Views and practices of reluctant teachers', *Prospect*, vol 17, no 3, pp 3-18

McNamara, T, 1996, *Measuring Second Language Performance*, Longman, London/New York

Munro, MJ and Derwing, TM, 1995a, 'Foreign accent, comprehensibility, and intelligibility in the speech of second language learners', *Language Learning*, vol 45, no 1, pp 73-97

Munro, MJ and Derwing, TM, 1995b, 'Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech', *Language & Speech*, vol 38, no 3, pp 289-306

Munro, MJ and Derwing, TM, 2001, 'Modeling perceptions of the accentedness and comprehensibility of L2 speech: the role of speaking rate', *Studies in Second Language Acquisition*, vol 23, no 4, pp 451-468

Munro, MJ, Derwing, TM and Morton, SL, 2006, 'The mutual intelligibility of L2 speech', *Studies in Second Language Acquisition*, vol 28, no 1, pp 111-131

Orr, M, 2002, 'The FCE Speaking test: using rater reports to help interpret test scores', in *System*, vol 30, no 2, pp 143-154

Schmid, PM and Yeni-Komshian, GH, 1999, 'The effects of speaker accent and target predictability on perception of mispronunciations', *Journal of Speech, Language, and Hearing Research*, vol 42, no 1, pp 56-64

Segalowitz, N, 2010, *Cognitive bases of second language fluency*, Routledge, London

Taylor, L and Jones, N, 2001, 'Revising the IELTS Speaking Test', *Research Notes*, vol 4, pp 9-12

Zielinski, BW, 2008, 'The listener: no longer the silent partner in reduced intelligibility', *System*, vol 36, no 1, pp 68-84

APPENDIX 1: QUESTIONNAIRES

Questionnaire A

(Note: In order to conserve space, the lines provided for answers have not been included in this version.)

Thank you for agreeing to participate in this study. We are interested in your views and experiences of assessing pronunciation as an examiner using the new IELTS Pronunciation scale. All your responses are strictly confidential.

How many years have you been an IELTS examiner? _____ years.

How many years have you been teaching ESL / EFL? _____ years.

What languages do you speak?

What language did you speak when you were growing up?

What language do you speak at home now?

In which countries have you lived and for how long?

What qualification(s) do you have? (Tick one or more)

- ❑ Diploma in Education (TESOL method)
- ❑ Graduate Diploma in _____
- ❑ CELTA
- ❑ Bachelor of Education (TESOL)
- ❑ Other (please specify) _____
- ❑ Graduate Certificate in _____
- ❑ Masters in _____
- ❑ DELTA
- ❑ Bachelor of Arts (Major: _____)
- ❑ Other (please specify) _____

1. How easy have you found the descriptors to use on the following IELTS Speaking test scales?

	very easy				**very hard**
	1	2	3	4	5
Fluency and Coherence	☐	☐	☐	☐	☐
Lexical resource	☐	☐	☐	☐	☐
Grammatical range and accuracy	☐	☐	☐	☐	☐
Pronunciation	☐	☐	☐	☐	☐

Give reasons for your answer

2. How confident do you feel about the accuracy of your rating on the following scales?

	not very confident				**very confident**
	1	2	3	4	5
Fluency and Coherence	☐	☐	☐	☐	☐
Lexical resource	☐	☐	☐	☐	☐
Grammatical range and accuracy	☐	☐	☐	☐	☐
Pronunciation	☐	☐	☐	☐	☐

Give reasons for your answer

3. How easy do you find it to:

	very easy				**very hard**
	1	2	3	4	5
(a) Use the increased number of Band levels on the Pronunciation scale?	☐	☐	☐	☐	☐
(b) Distinguish between Band levels for pronunciation?	☐	☐	☐	☐	☐
(c) Understand the descriptors	☐	☐	☐	☐	☐

4a. How confident do you feel when you are judging the following features of a candidate's speech?

	not very confident				**very confident**
	1	2	3	4	5
Sounds	☐	☐	☐	☐	☐
Rhythm	☐	☐	☐	☐	☐
Stress (word level)	☐	☐	☐	☐	☐
Stress (sentence level)	☐	☐	☐	☐	☐
Intonation	☐	☐	☐	☐	☐
Chunking (pausing)	☐	☐	☐	☐	☐
Speech rate	☐	☐	☐	☐	☐
Intelligibility	☐	☐	☐	☐	☐
Listener strain	☐	☐	☐	☐	☐
Accent	☐	☐	☐	☐	☐

4b. Which of these features of spoken language do you think are most important when you are awarding a pronunciation score? Please rank them if appropriate.

5. When you re-certified on the new Pronunciation scale, did you have:

a group session with an IELTS trainer or individual self access? (Underline your answer)

6. How well do you feel the training prepare you to examine using the revised Pronunciation scale?

not very well				**very well**
1	2	3	4	5
☐	☐	☐	☐	☐

Please give reasons for your answer.

7. If you are familiar with previous Band scale, which scale do you prefer? Underline your preferred answer.

The previous 4 band scale or The revised 9 band scale?

Why?

8. Do you have any other comments on the revised Pronunciation scale?

Questionnaire B

(Note: to conserve space, the rating scales for all speakers are not included in this version.)

Version :_____ **Participant number: _____**

Rating Task 1

You will hear 12 recordings of Part 3 of the IELTS Speaking Test, the 4-5 minute two way discussion.

For each speaker, listen to the recording, refer to the scales as you would when examining and then write **your IELTS Pronunciation Band score** in the space provided. You may listen to sections of the recording again as you make your decision as in the IELTS test situation. Then circle the number that best represents how confident you feel in the accuracy of your rating.

Speaker 1

IELTS Pronunciation Band score: _____

How confident are you that this rating is accurate? *Not at all confident* *Very Confident*

1 2 3 4 5

Speaker 2

IELTS Pronunciation Band score: _____

How confident are you that this rating is accurate? *Not at all confident* *Very Confident*

1 2 3 4 5

Speaker 3

IELTS Pronunciation Band score: _____

How confident are you that this rating is accurate? *Not at all confident* *Very Confident*

1 2 3 4 5

Questionnaire C

(Note: In order to conserve space, the spaces provided for answers have not been included in this version.)

Participant no: _____

Level distinctions

1. How easy did you find it to distinguish between Pronunciation Band levels for these candidates?

very easy 1 2 3 4 5 **very hard**

Give details

2. Were there any Pronunciation Bands you found it difficult to choose between?
Yes/ No If yes, which ones and why?

3. Which statement best fits your understanding of the rationale for awarding a Pronunciation Band 5?
Circle your answer.

1. The candidate displays all the features of 4 and most of the positive features of 6.
2. The candidate displays all of the features of 4 and all but one of the positive features of 6.
3. The candidate appears to be mid-way between a 4 and a 6.

Comments:

The Pronunciation Descriptors

4. When you were assessing Pronunciation which part(s) of the descriptor did you generally find yourself paying most attention to?

5. Do you think the descriptors are about the right length or would you prefer them to be shorter/longer? Please elaborate.

6. Do you think the descriptors cover features of pronunciation that can be readily assessed in the testing situation? Yes/no. Please elaborate.

7. Are there aspects of pronunciation you think are important that are not mentioned in the descriptors? If so, please note them below.

The Rating Process

8. Which part of the test is most useful to you when making a judgement about pronunciation? Please circle the best answer:

Part 1 (Introduction and interview)
Part 2 (Individual Long turn)
Part 3 (Two way discussion)

Why?

9. How is your final Pronunciation rating achieved? How do you work towards it? At what point do you finalise your Pronunciation rating?

Comments

10. What do you like about the new Pronunciation scale?

11. What don't you like about the revised Pronunciation scale?

12. In your opinion, how could the Pronunciation scale be improved?

13. Any other comments?

APPENDIX 2: CODING CATEGORIES FOR VP COMMENTS

Accent
The word *accent* is used by the VP examiner.

Affective comments
Comments that reflect on what the candidate might be feeling or thinking.

Chunking
The word *chunking* is used by the VP examiner or the VP examiner indicates that the candidate pauses in the right place.

Clarity
The word *clarity* is used by the VP examiner. Includes comments related to how *clear* a candidate's speech is.

Connected speech level
Comments on anything above the word level. Includes stress at sentence level.

Effort required to understand candidate
The degree of effort required of the listener to understand the candidate. Includes comments related to how hard a candidate is to understand.

Features contributing directly to decision on band level assigned
Any connection between a feature of pronunciation and the band level assigned or the decision making process of assigning a band level.

Non-Pronunciation scale comments
Comments on features and/ or use of terms that are not included in the band descriptors or key indicators for the revised IELTS Pronunciation scale.

Intelligibility
The VP examiner either (1) uses the word *intelligibility*, (2) comments she can't understand what the candidate is saying, or (3) indicates that intelligibility (word recognition) has been affected - e.g., a particular feature has contributed to making what the candidate said sounding like something else, or a particular feature makes it easy to recognise the words a candidate says.

Intonation
The word *intonation* is used or the VP examiner's comments are related to tone or pitch variation.

Linking
Comments related to linking words together – related to phonemes rather than rhythm.

Negative comment
Comments about something the candidate is doing wrong.

Phonemes
The word *phoneme* is used or comments relate to sounds, consonants or vowels.

Positive comment
Comments about something the candidate is doing right or well.

Rhythm
The word *rhythm* is used or comments relate to timing (eg, stress timing, syllable timing) or linking of words in connected speech.

Speech rate
Comments related to the rate of speech.

Stress
Comments related to stress in words or stress of words in sentences.

Stress at word level
Comments related to stress patterns in individual words.

Stress in connected speech
Comments related to the stress pattern across sections of connected speech.

Word level
Comments related to individual words.

APPENDIX 3: STATISTICAL ANALYSIS

Analysis for Table 5: Ease of use of descriptors: Paired-sample t-test values

	Fluency & Coherence *M* = 2.41,*SD* = 1.047	Lexical Resource *M* = 2.26, *SD* = 0.903	Grammatical Range & Accuracy *M* = 2.59, *SD* = 0.971
Pronunciation *M* = 2.81, *SD* = 0.962	1.954	2.749*	1.100

Note: df =26; * p<.05

Analysis for Table 7: Confidence judging features of pronunciation: Paired-sample t-test values

Concrete features	Global judgements		
	Intelligibility *M* = 4.19, *SD* = 1.001	Listener effort *M* = 4.07, *SD* = 1.072	Accent *M* = 3.96, *SD* = 0.980
Sounds *M* = 3.78, *SD* = 0.934	3.328*	2.530*	1.095
Rhythm *M* = 3.52, *SD* = 0.849	4.416*	3.238*	2.590*
Word stress *M* = 3.74, *SD* = 1.095	2.884*	1.975	1.100
Sentence stress *M* = 3.67, *SD* = 1.074	3.578*	2.383*	1.551
Intonation *M* = 3.67, *SD* = 1.038	3.358*	2.383*	1.442
Chunking *M* = 3.74, *SD* = 1.023	3.075*	1.975	1.363
Speech rate *M* = 3.96, *SD* = 0.940	1.442	0.721	means are the same

Note: df =26; * p<.05