Re-examining the validation of a holistic speaking scale: the (non-) applicability of pronunciation descriptors.

Philip Horne
British Council

Research Context I

- Aptis was launched by British Council in 2012.
- Aptis General is a proficiency test of English, with components targeting all four skills.
- The speaking and writing components are graded by accredited and trained raters using a holistic scale.
- The speaking holistic scales include descriptors targeting : fluency, coherence, grammar, lexis, and pronunciation.
- The levels are CEFR-linked and benchmarked.

Research Context II

- Within the CEFR, pronunciation was conspicuous only due to its apparent neglect (Harding, 2017).
- In recognition of these limitations, the CEFR produced an updated version of the scale (Council of Europe, 2018).

Literature Review

- **Derwing & Munro, (1999; 2009); Isaacs & Trofimovich, (2012)**

Accentedness, comprehensibility, intelligibility. How do we define and distinguish these constructs, and how do we operationalise them for the purposes of reliable rating?

- **Isaacs et al (2015)**

Studied the current IELTS scales, and found that the lack of precision (especially at higher levels) was detrimental to assessment.

- **Harding (2017)**

Carried out a study focusing upon rater interpretation of the original Phonological Control Scale and found it was beset with issues.

- **Lumley (2002)**

Focused on writing scales but studied how raters turn to their own understanding of the construct when an existing rating scale is found to lack logic.

Original CEFR Phonological Control Scale

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

Council of Europe (2001, p.117).

Updated CEFR Phonological Control Scale

	Descriptor
6 C2	Can employ the full range of phonological features in the target language with a high level of control – including prosodic features such as word and sentence stress, rhythm and intonation. The finer points of his/her message are clear and precise. Intelligibility is not affected in any way by features of accent that may be retained from other language(s).
5 C1	Can employ the full range of phonological features in the TL with sufficient control to ensure intelligibility throughout. Can articulate virtually all the sounds of the TL. Some features of accent retained from other language(s) may be noticeable, but they do not affect intelligibility at all.
4 B2	Can generally use appropriate intonation, place stress correctly. Can articulate individual sounds clearly. Accent tends to be influenced by other language(s) he/she speaks but has little or no effect on intelligibility.
3 B1	Pronunciation is generally intelligible. Can approximate intonation and stress at both utterance and word levels. Accent is usually influenced by other language(s) he/she speaks.
2 A2	Pronunciation is generally clear enough to be understood, but conversational partners will need to ask for repetition from time to time. A strong influence from other language(s) he/she speaks on stress, rhythm and intonation may affect intelligibility, requiring collaboration from interlocutors. Pronunciation of familiar words is clear.
1 A1	Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by interlocutors used to dealing with speakers of the language group concerned. Can reproduce correctly a limited range of sounds as well as the stress on simple, familiar words and phrases.

Research Context III

- Is pronunciation in greater than usual danger of conflation when applying a global score?

- Does the update to the CEFR Phonological Control Scale present a renewed need for investigation into the benchmarking between the CEFR and the CEFR-linked descriptors in the Aptis holistic scale?

Existing Aptis Task Four Scale

5 **C1**	**Response addresses all three questions and is well-structured.** • **Uses a range of complex grammar constructions accurately. Some minor errors occur but do not impede understanding.** • **Uses a range of vocabulary to discuss the topics required by the task. Some awkward usage or slightly inappropriate lexical choices.** • **Pronunciation is clearly intelligible.** • **Backtracking and reformulations do not fully interrupt the flow of speech.** • **A range of cohesive devices are used to clearly indicate the links between ideas.**
4 **B2.2**	Responses to all three questions are on topic and show the following features • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. • **Pronunciation is intelligible. Mispronunciations do not put a strain on the listener or lead to misunderstanding** • Some pausing while searching for vocabulary but this does not put a strain on the listener. • A limited number of cohesive devices are used to indicate the links between ideas.

Evaluating the Existing Aptis Pronunciation Descriptors

- Rather vague, especially in comparison with their counterparts.

- The detail is focused more upon negative aspects of the performance.

- No precision at higher levels – rely upon raters understanding and interpreting correctly (for example) the difference between “intelligible” and “clearly intelligible”.

Research Questions

- To what extent is there correlation between the scores (CEFR levels) awarded using the Aptis Task Four holistic scale and the updated CEFR Phonological Control Scale?

- Which features of speech factor into rater decision-making when interpreting pronunciation descriptors in the Aptis Task Four holistic speaking scale?

- Do the same features of speech factor into rater decision-making when applying more level-specific pronunciation descriptors in the updated CEFR Phonological Control Scale?

Methodology

Overview

- Three phases: two rating sessions and a series of paired interviews with all the raters involved.

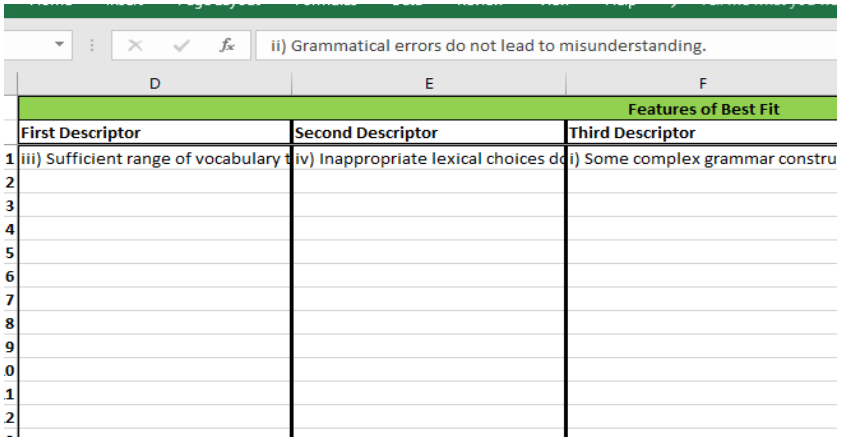
- Mixed-Methods

Task Four and CEFR Alignment

Numeric Score	Aptis Task Four	Phonological Control Scale
6	C2	C2
5	C1	C1
4	B2.2	B2
3	B2.1	B1
2	B1.2	A2
1	B1.1	A1
0	A1/A2	-

Rating Session One

- 6 raters mark (under authentic conditions) 42 speaking samples (Aptis Task Four)
- They then select from a drop down list the five descriptors which informed their decision (in rank order).



Task Four Rating Scale for Rating Session One

5 **C1**	**Response addresses all three questions and is well-structured.** **i) Uses a range of complex grammar constructions accurately.** **ii) Some minor grammatical errors occur but do not impede understanding.** **iii) Uses a range of vocabulary to discuss the topics required by the task.** **iv) Some awkward usage or slightly inappropriate lexical choices.** **v) Pronunciation is clearly intelligible.** **vi) Backtracking and reformulations do not fully interrupt the flow of speech.** **vii) A range of cohesive devices are used to clearly indicate the links between ideas.**
4 **B2.2**	Responses to all three questions are on topic and show the following features. i) Some complex grammar constructions used accurately. ii) Grammatical errors do not lead to misunderstanding. iii) Sufficient range of vocabulary to discuss the topics required by the task. iv) Inappropriate lexical choices do not lead to misunderstanding. v) Pronunciation is intelligible. vi) Mispronunciations do not put a strain on the listener or lead to misunderstanding. vii) Some pausing while searching for vocabulary but this does not put a strain on the listener. viii) A limited number of cohesive devices are used to indicate the links between ideas.

Recoded Descriptors in the Task Four Scale

In the original Task Four Scale at each level…	Recoded as…
First Fluency Descriptor	FCa
Second Fluency Descriptor	FCb
Third Fluency Descriptor	FCc
First Lexical Descriptor	LRa
Second Lexical Descriptor	LRb
First Grammar Descriptor	GRAa
Second Grammar Descriptor	GRAb
First Pronunciation Descriptor	PROa
Second Pronunciation Descriptor	PROb

Rating Session Two

- The same six raters mark the same 42 samples of speech (in a randomised order).
- This time they apply the updated Phonological Control Scale (focusing exclusively upon pronunciation-related facets of speech).
- They then select from a drop-down list the descriptors which were most relevant to their decision (in rank order).

Recoded Descriptors in the Phonological Control Scale

	Descriptor
6 **C2**	i) Can employ the full range of phonological features in the target language with a high level of control – including prosodic features such as word and sentence stress, rhythm and intonation. ii) The finer points of his/her message are clear and precise. iii) Intelligibility is not affected in any way by features of accent that may be retained from other language(s).
5 **C1**	i) Can employ the full range of phonological features in the TL with sufficient control to ensure intelligibility throughout. ii) Can articulate virtually all the sounds of the TL. iii) Some features of accent retained from other language(s) may be noticeable, but they do not affect intelligibility at all.
4 **B2**	i) Can generally use appropriate intonation, place stress correctly. ii) Can articulate individual sounds clearly. iii) Accent tends to be influenced by other language(s) he/she speaks but has little or no effect on intelligibility.

Recoded Descriptors in the Phonological Control Scale

In the original Phonological Control Scale at each level…	Recoded as…
Intelligibility-focused Descriptor	Intelligibility
Accent-focused Descriptor	Accent
Phonological Control-focused Descriptor	PhonC

Interviews

- Three interviews are conducted (six raters divided into pairs).

- Based upon emergent trends from statistical analysis (as well as pre-determined questions such as understanding of the intelligibility construct), an interview schedule is devised,

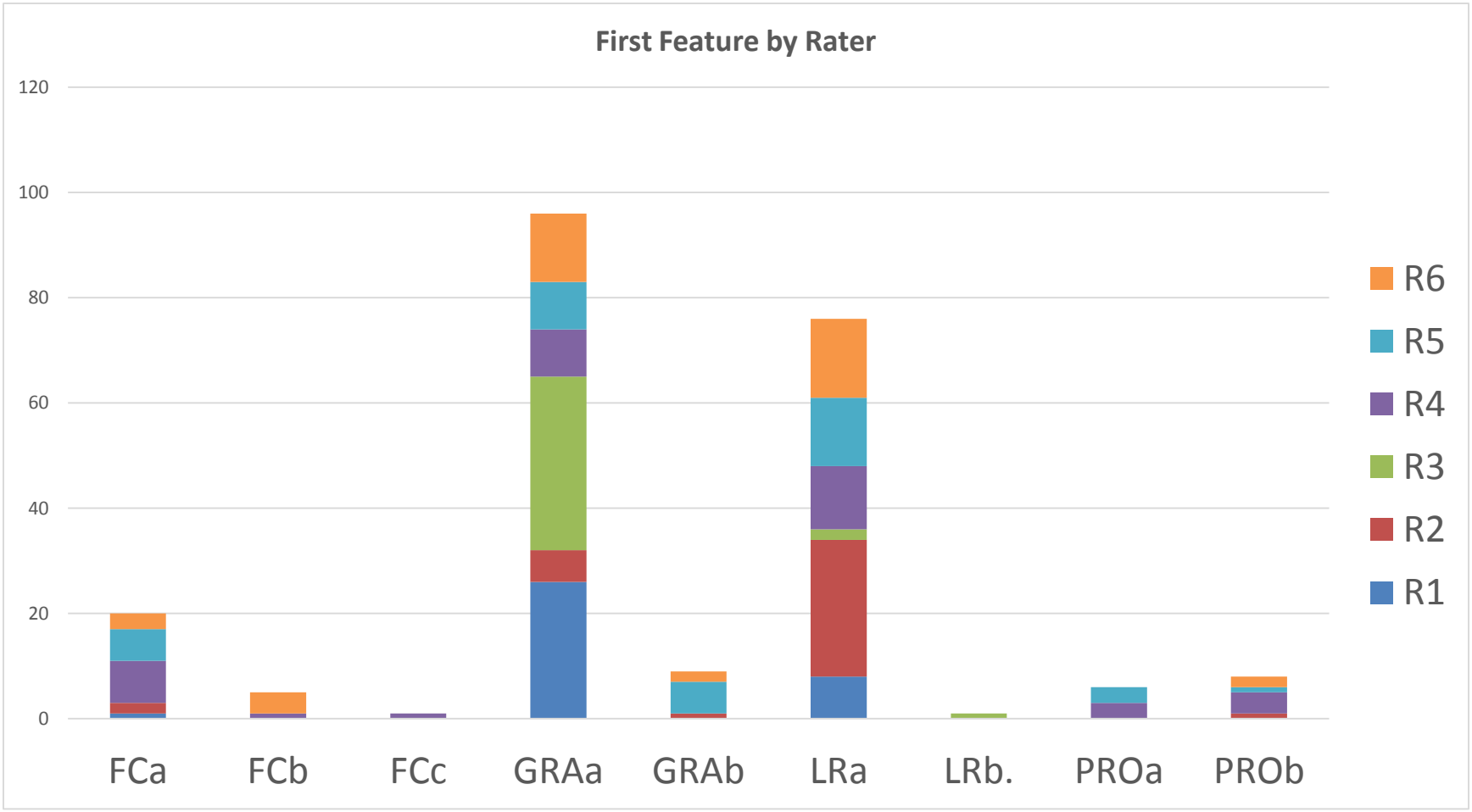
Results: Rating Session One

Inter-Rater Correlation: Rating Session One

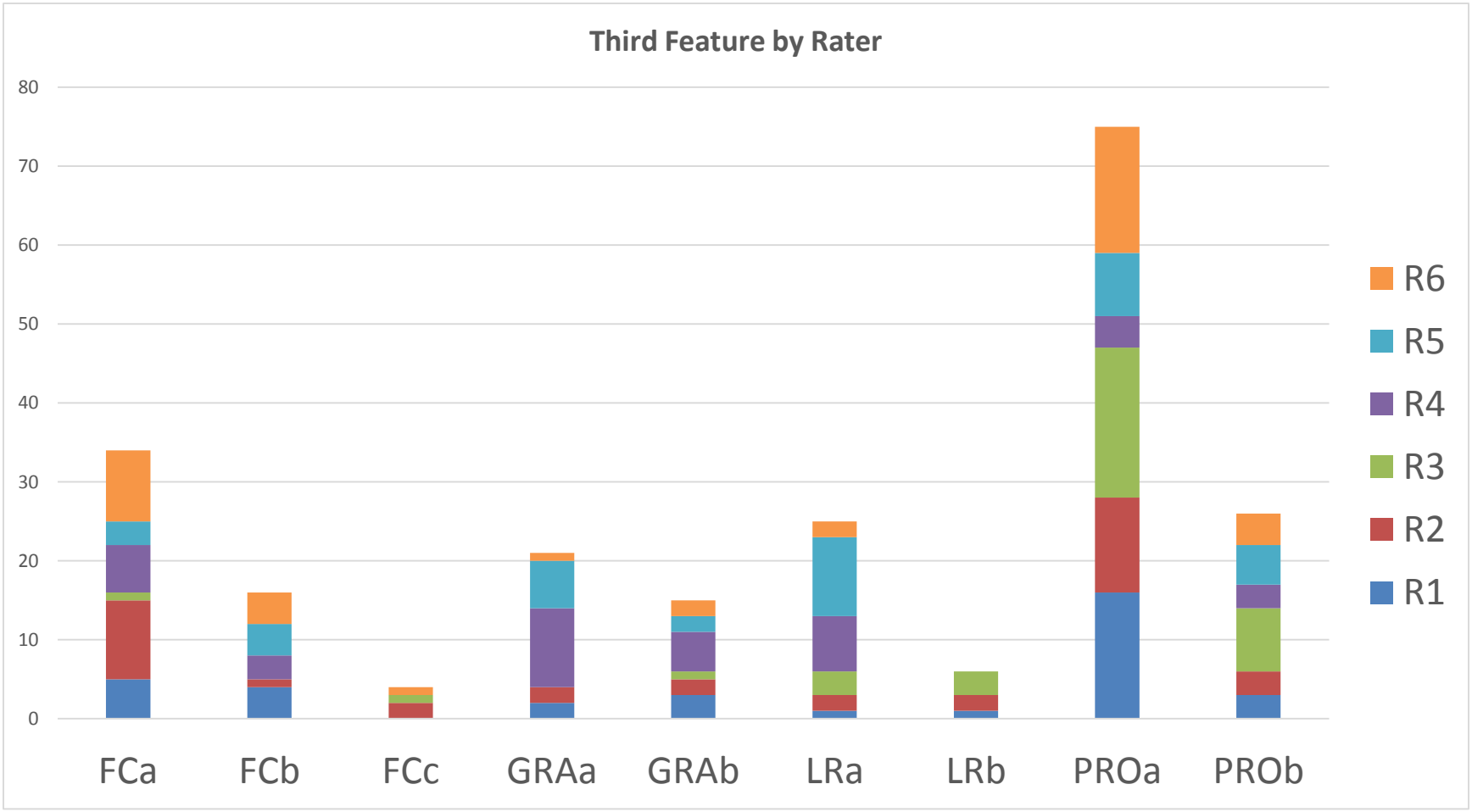
	R1	R2	R3	R4	R5	R6
R1	1					
R2	.87	1				
R3	.85	.89	1			
R4	.79	.83	.84	1		
R5	.82	.85	.85	.93	1	
R6	.79	.79	.83	.73	.73	1

All figures are significant at the $p < .001$ level.

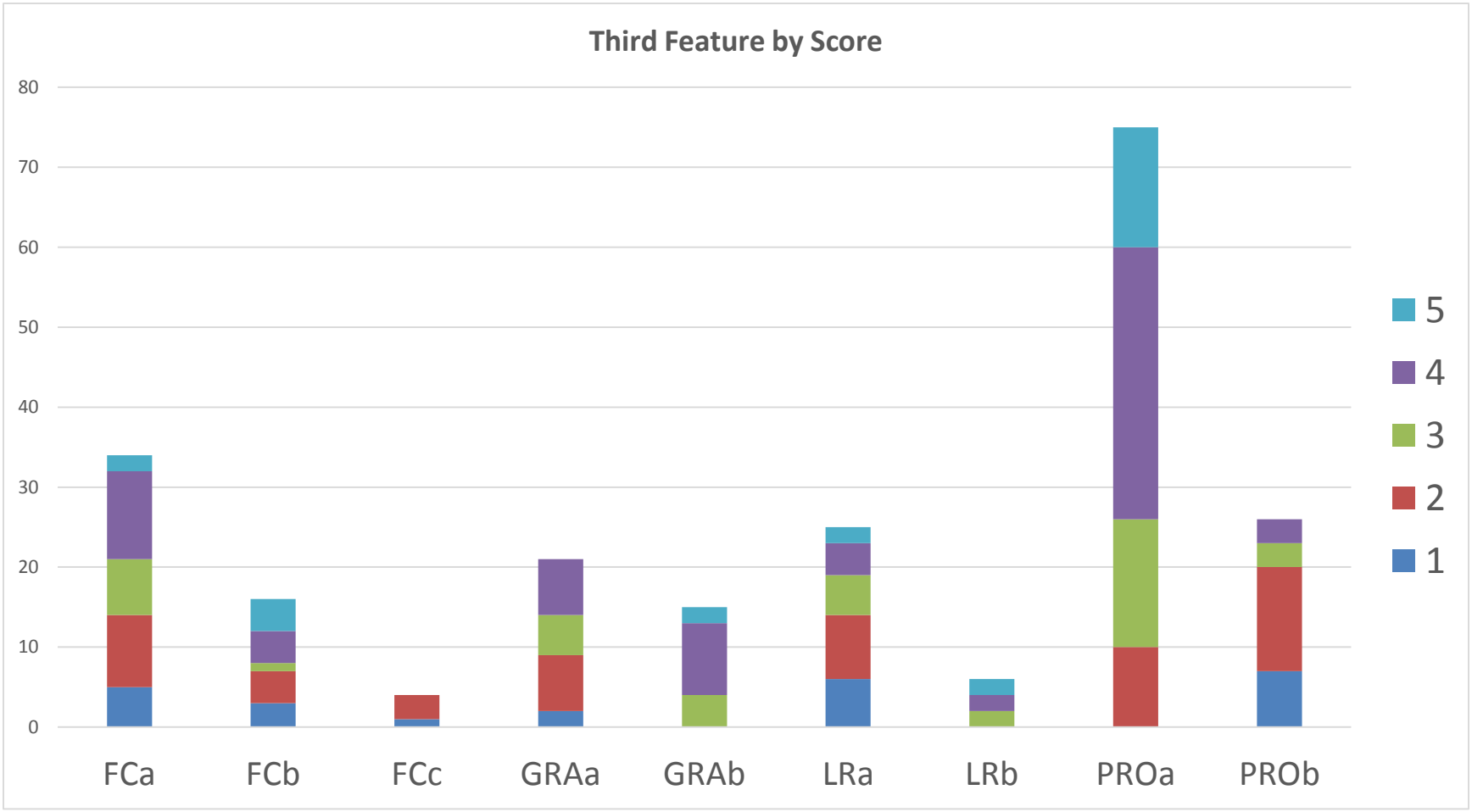
Results: Rating Session One



Results: Rating Session One



Results: Rating Session One



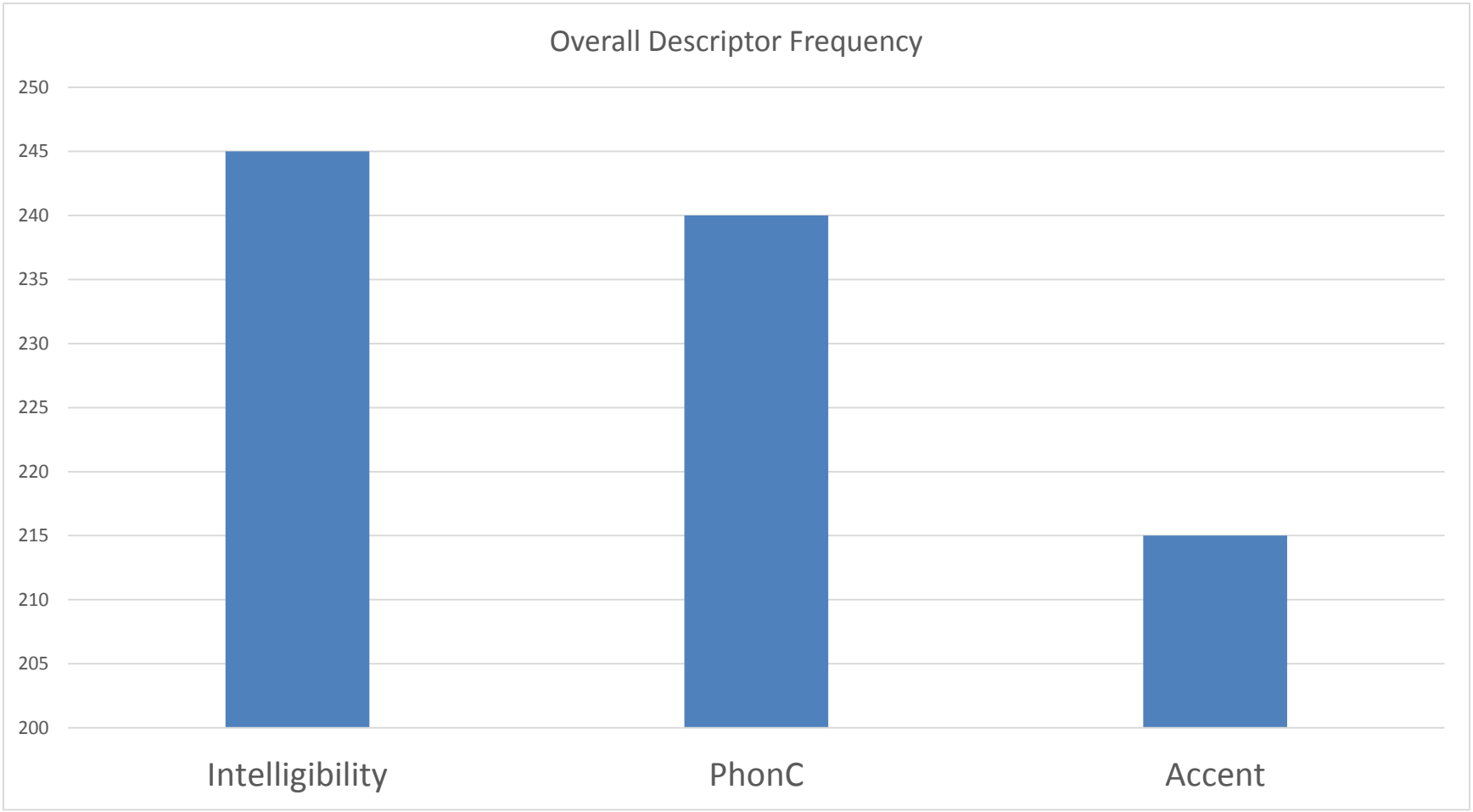
Results: Rating Session Two

Inter-Rater Correlation: Rating Session Two

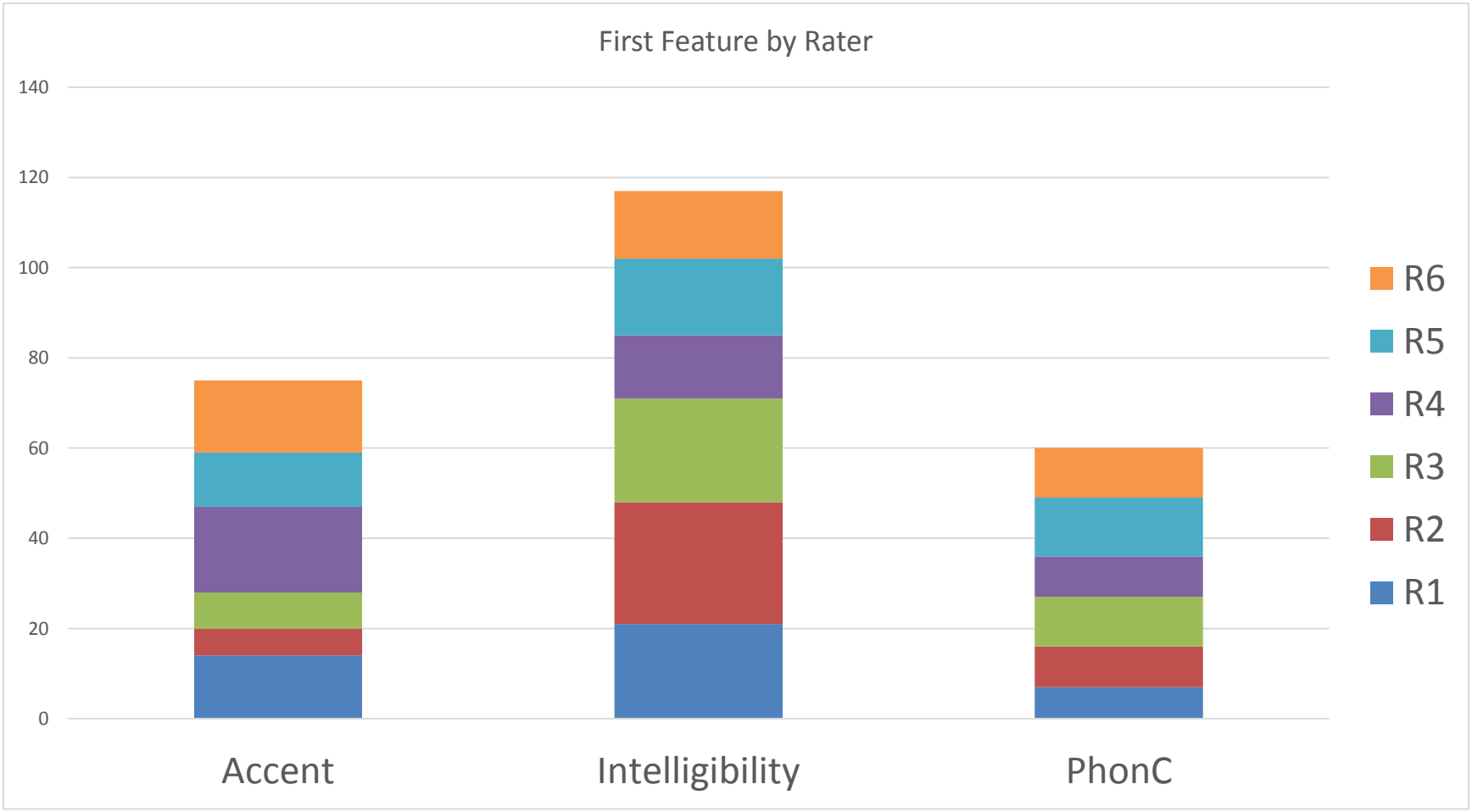
	R1	R2	R3	R4	R5	R6
R1	1					
R2	.67	1				
R3	.67	.69	1			
R4	.52	.73	.71	1		
R5	.75	.76	.74	.70	1	
R6	.74	.74	.70	.65	.71	1

Table 4-2*: Inter-Rater Correlation in Rating Session Two*
All figures are significant at the p <.001 level.

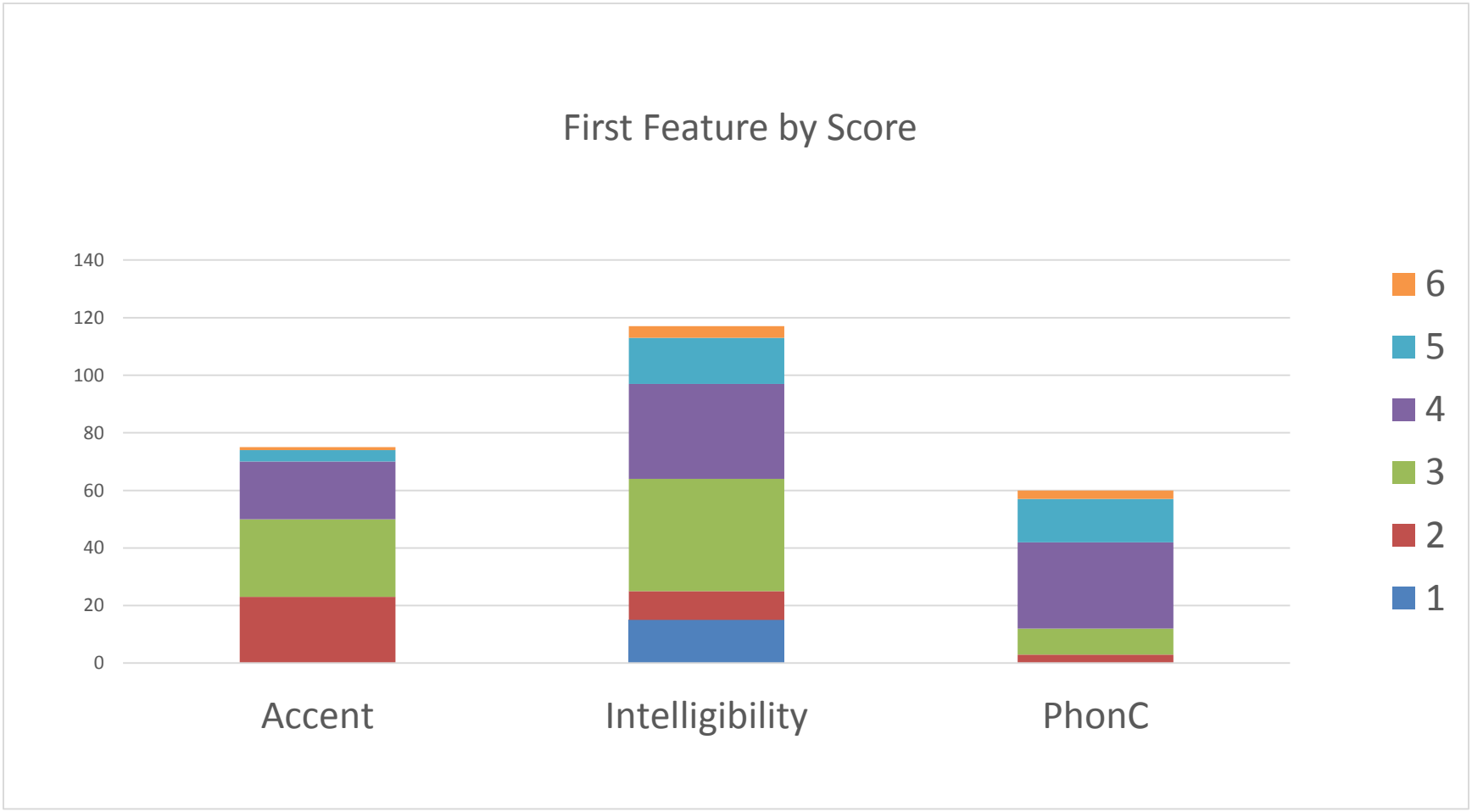
Results: Rating Session Two



Results: Rating Session Two



Results: Rating Session Two



Results: Rating Session Comparison

Results: Rating Session Comparison

Table 4-3*: Comparative Descriptive Mean Scores*

	Rating Session One	Rating Session Two
Mean	2.84	3.44
Standard Error	.23	.15
Median	2.92	3.59
Mode	4	3.67
Standard Deviation	1.42	1.00

Results: Rating Session Comparison

Table 4-4: *Rating Session One and Rating Session Two Intra-Rater Correlation*
All figures are significant at the p <.001 level

	r_s
R1	.78
R2	.78
R3	.79
R4	.71
R5	.77
R6	.86

Results: Rating Session Comparison

Table 4-5: *Recoded Scores from the Task Four Scale and the Phonological Control Scale*

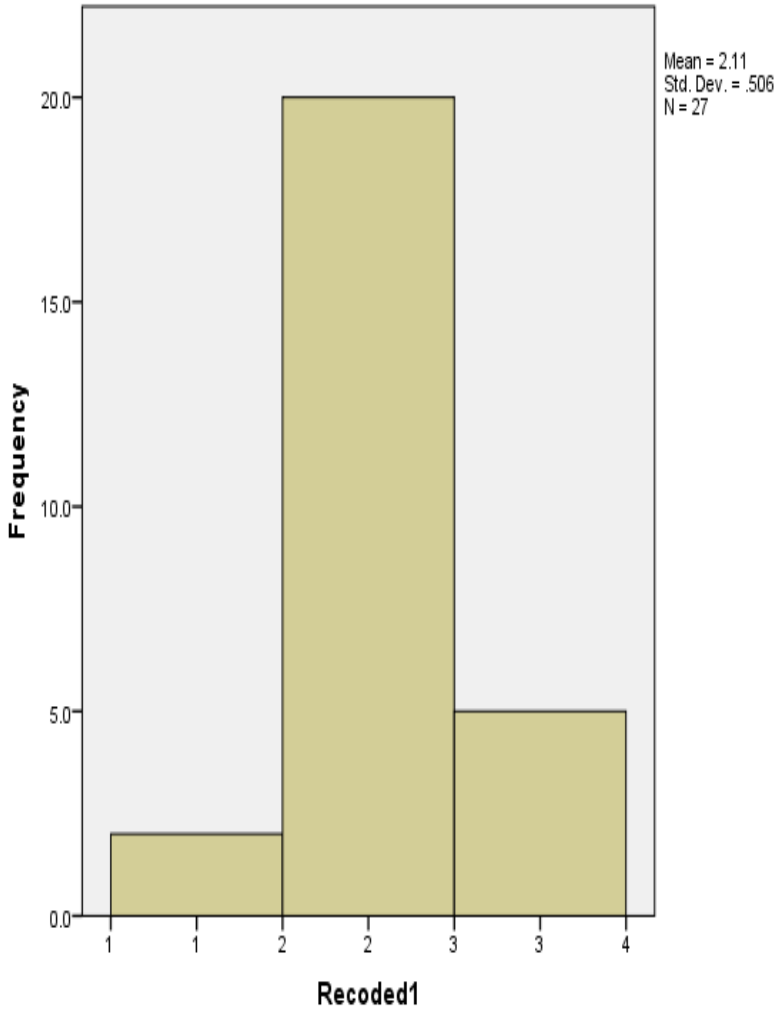
Original Task Four Score	Original Phonological Control Scale Score	Recoded as…
C2 (6)	C2 (6)	4
C1 (5)	C1 (5)	3
B2.2/B2.1 (4/3)	B2 (4)	2
B1.2/B1.1 (1/2)	B1 (3)	1

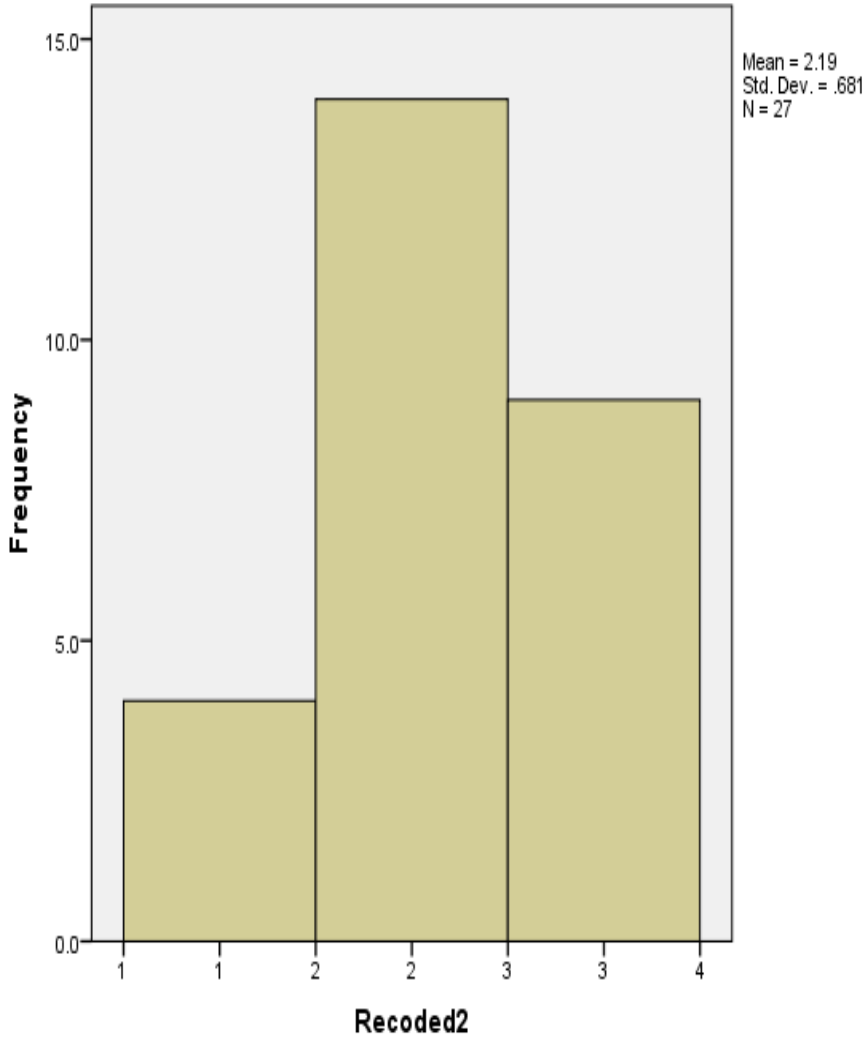
Results: Rating Session Comparison

Table 4-6*: Descriptive Statistics based upon Recoded Scores*

	Rating Session One (Recoded One)	Rating Session Two (Recoded Two)
Mean	2.11	2.19
Standard Error	.10	.13
Median	2	2
Mode	2	2
Standard Deviation	.51	.68

Results: Rating Session Comparison





Recoded1 * Recoded2 Crosstabulation

			Recoded2			Total
			1	2	3	
Recoded1	1	Count	1	1	0	2
		Expected Count	.3	1.0	.7	2.0
		Residual	.7	.0	-.7	
		Std. Residual	1.3	.0	-.8	
	2	Count	3	11	6	20
		Expecte Count	3.0	10.4	6.7	20.0
		Residual	.0	.6	-.7	
		Std. Residual	.0	.2	-.3	
	3	Count	0	2	3	5
		Expected Count	.7	2.6	1.7	5.0
		Residual	-.7	-.6	1.3	
		Std. Residual	-.9	-.4	1.0	
Total		Count	4	14	9	27
		Expected Count	4.0	14.0	9.0	27.0

Results: Interviews

Interviews

- Intelligibility is a superficial judgement.
- Distinguishing between “clearly intelligible and “intelligible”. introduces self-operationalization and subjectivity.
- Pronunciation only becomes a focus when it is an issue.
- There is desire for more precision, especially at higher levels.
- This would help discriminate better between C-level candidates.
- Participants also emphasized the importance of retaining the holistic nature of rating and not to go too far in the other direction – i.e. over-emphasis any one criteria.

Conclusions

Limitations and Follow-up

- Small sample (especially when collapsing the dataset to focus on aligned levels)
- Limited number of raters
- Raters did not have training or standardization in applying Phonological Control Scale.

- Follow-up with robust statistical analysis and multi-faceted Rasch
- Incorporate revised pronunciation descriptors into the existing Task Four Aptis Scale (together with a control group).

Conclusions

- If pronunciation descriptors are to be included in holistic scales, the precise features of language being considered need clear definition.
- These must include both positive and negative features of the performance.
- Intelligibility is overly broad and alone is only a superficial judgement.
- Only including negatively-worded descriptors results in pronunciation becoming less applicable at higher levels.

References

- Derwing, T. M. & Munro, M.J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379-397.

- Derwing, T. M. & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42(4)*, 476-490.

- Harding, L. (2017). What do raters need in a pronunciation scale: A user's perspective. In: T. Isaacs and P. Trofimovich (Eds). *Second language pronunciation assessment: Interdisciplinary perspectives*. UK: CPI Book Groups.

- Isaacs, T., Trofimovich, P., Yu, G. and Chereau, B.M. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. IELTS Research Reports Online 4, 1–48.

- Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing 19(3)*, 246-276.

THANK YOU