ISSN 2201-2982
2021/5

IELTS Research Reports Online Series

VEO IELTS PROJECT REPORT:

Which specific features of candidate talk do examiners orient to when taking scoring decisions?

Paul Seedhouse and Müge Satar

VEO IELTS PROJECT REPORT: Which specific features of candidate talk do examiners orient to when taking scoring decisions?

The research investigated which specific features of candidate talk IELTS Speaking Test (IST) examiners orient to when taking scoring decisions. We also researched whether the use of the scoring scheme and customised app potentially adds any value to IST examiner development.

Funding

This research was funded by the IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia. Grant awarded 2018.

Publishing details

Published by the IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia © 2021.

How to cite this article

Seedhouse, P., & Satar, M. (2021). VEO IELTS Project Report: Which specific features of candidate talk do examiners orient to when taking scoring decisions? *IELTS Research Reports Online Series, No. 5*. British Council, Cambridge Assessment English and IDP: IELTS Australia.
Available at https://www.ielts.org/teaching-and-research/research-reports

Introduction

This study by Seedhouse and Satar was conducted with support from the IELTS partners (British Council, IDP: IELTS Australia and Cambridge Assessment English), as part of the IELTS joint-funded research program. Research funded by the British Council and IDP: IELTS Australia under this program complement those conducted or commissioned by Cambridge Assessment English, and together inform the ongoing validation and improvement of IELTS.

A significant body of research has been produced since the joint-funded research program started in 1995, with over 130 empirical studies receiving grant funding. After undergoing a process of peer review and revision, many of the studies have been published in academic journals, in several IELTS-focused volumes in the *Studies in Language Testing* series (http://www.cambridgeenglish.org/silt), and in the *IELTS Research Reports*. Since 2012, to facilitate timely access, individual research reports have been made available on the IELTS website immediately after completing the peer review and revision process.

This research investigated which specific features of candidate talk IELTS Speaking Test Examiners orient to when making scoring decisions. Part of the research was also to explore whether the use of a scoring scheme and a customised app had the potential to add value to Examiner development. This work contributes to the very important body of research into Examiner behaviour as part of the rating process. Examiners are critical stakeholders in the testing and assessment field and as such, Examiner training, monitoring and development involve rigorous and thorough processes and procedures. Studies such as this are welcomed by the IELTS Partners as they seek to investigate possible new avenues for Examiner development in an ever-changing environment.

The researchers collected the data in two ways. Firstly, by asking Examiners to use the VEO app together with the customised scoring system. Secondly, qualitative data was collected through Examiner focus groups. Although the examiner cohort was small, small-scale studies can provide highly valuable insights when we remain cognisant of these limitations, and the broader conclusions to be drawn from them.

The findings of the study suggest that the use of the customised IELTS scoring system for Speaking and the VEO app has the potential to add value to Examiner development, particularly for re-certification. This was corroborated by all four Examiners during the focus groups. However, the use of these customised resources does not necessarily contribute to the process or accuracy of the rating process to the rating process. Examiners did not record all of the cognitive processes they went through, practically it was not possible. Instead they tended to mark cumulative scores when they noticed a pattern several times and were confident of a judgement. The graphical representation of findings provides Examiners and other stakeholders interesting data to consult when developing Examiner training.

This study has explored and suggested an interesting technological option to support Examiner development. Detecting specific features of candidate talk that trigger Examiner decisions is a complex task. At a time when test developers are looking for useful technological solutions for different stages of the testing process, this research has shed light on a possibility for Examiner development.

Mina Patel
Assessment Research Manager
British Council

Dr Tony Clark
Senior Research Manager
Cambridge Assessment English

VEO IELTS PROJECT REPORT: Which specific features of candidate talk do examiners orient to when taking scoring decisions

Abstract

We used an app to develop a new way of finding out what examiners notice when they give grades to students in speaking tests of English. This means we can better understand how examiners give marks, which can help with training examiners.

The research investigated which specific features of candidate talk IELTS Speaking Test (IST) examiners orient to when taking scoring decisions. We also researched whether the use of the scoring scheme and customised app potentially adds any value to IST examiner development.

This was enabled by the development of an IST scoring scheme for an app (VEO) which creates a recording of when exactly in the test the examiners have noticed specific features of candidate talk and taken specific decisions on scoring as a result. Each of four IELTS examiners independently viewed two test videos using the app and scored it using the scoring tags. We then conducted individual stimulated recall interviews with the examiners involved.

We found that the use of the customised IELTS scoring scheme and VEO app illuminates the IST rating process and potentially adds significant value to IST examiner development, specifically the re-certification process. When examiners assign higher scores, they focus on positive evidence, whereas with lower scores, they focus on negative evidence. Fluency & coherence scores are mostly assigned cumulatively. Grammar scores can be more easily tagged in relation to specific features. Examiners notice idioms and reward their use with a high mark, even if not delivered perfectly. Examiners form hypotheses as they listen to the candidate talk, then look for evidence that will confirm or reject these hypotheses. Examiners make scoring decisions in a cumulative way, rather than orientating towards single instances. Pronunciation issues may influence examiner decisions in relation to other criteria.

Authors' biodata

Paul Seedhouse

Paul Seedhouse is Professor of Educational and Applied Linguistics at Newcastle University, UK. Working with colleagues in Computer Science, he used three grants to build kitchens which use digital technology to teach users languages and cuisines simultaneously: https://linguacuisine.com. He has also had four grants to study interaction in the IELTS Speaking Test. The *IELTS Research Reports* on these projects are available on the IELTS website. His 2018 book, *The Discourse of the IELTS Speaking Test*, with Fumiyo Nakatsuhara is published by Cambridge University Press.

Müge Satar

Müge Satar is a Lecturer in Applied Linguistics and TESOL at Newcastle University, UK. She is interested in communicative and pedagogical aspects of multimodal interaction for online language learning and teaching, focusing on social presence, meaning-making, instruction-giving, and translanguaging. She is the co-editor of the *Journal of Virtual Exchange* and has publications in leading journals in the field, such as *Language Learning and Technology, ReCALL* and *System*.

Acknowledgements

Many thanks to Şaziye Taşdemir for her excellent work on the project in terms of data collection and editing the report.

Contents

1 Introduction 10

1.1 Literature review 10

1.2 The current study 12

1.3 Background information on the VEO app........ 13

1.4 How does the IELTS tagset work with the VEO app? 14

2 Research design and procedures........ 16

3 Findings 18

Definitions........ 19

PART 1 19

3.1 When do examiners make scoring decisions?........ 19

Graphical representation of the data 22

3.2 Are examiners consistent in their decisions on scores and decision times?........ 23

3.2.1 Inter-rater reliability for overall scores........ 25

3.2.2 Inter-rater reliability figures for decision times........ 26

3.3 Which salient features do examiners orient to?........ 27

3.3.1 Fluency and coherence........ 27

3.3.1.1 Fluency & coherence scores are mostly assigned cumulatively, and there is often a weak correlation between specific and overall/final scores. 27

3.3.1.2 For fluency and coherence, when examiners assign higher scores, they seem to focus on positive evidence. However, when they assign lower scores, they seem to focus on negative evidence. 28

3.3.2 Grammatical range and accuracy........ 29

3.3.2.1 For grammatical range and accuracy, examiners may take a decision on grammar at exactly the same time, but they may assign different scores as they focus on different features of candidate talk. 29

3.3.2.2 Grammar seems to be a criterion that can be more easily tagged in relation to specific features, especially compared with fluency and coherence. 30

3.3.3 Lexical resource 31

3.3.3.1 Examiners tend to notice idioms and reward their use with a high mark, even if they are not delivered quite perfectly. 31

3.3.3.2 'Rehearsed' or 'memorised' chunks may not receive as high a score as other 'sophisticated' idioms which are employed in a more context-sensitive way. 32

3.3.3.3 When the candidate is not able to use simple vocabulary, this might be noticed by the examiners, and marked with a lower score. 33

3.3.4 Pronunciation........ 34

3.3.4.1 Specific scores may relate to specific problems noticed for pronunciation, whereas cumulative scores may be higher. 34

3.3.4.2 When examiners assign higher scores for pronunciation, they tend to focus on positive aspects (i.e. what the candidate can do); and when they assign lower scores, they focus on negative aspects (i.e. what the candidate cannot do). This provides additional support for argument 3.3.1.2. 35

3.3.4.3 All examiners may notice the same salient point at exactly the same time. However, a single item may not affect the candidate's scores significantly as examiners look for sustained evidence. 35

3.3.4.4 For pronunciation, all examiners may assign the same scores at roughly the same time. However, examiners may notice different aspects of candidate talk and provide different descriptions of the pronunciation problem. 37

3.4 How do examiners decide when to give a score? 38

3.4.1 Examiners start forming hypotheses as they listen to the candidate talk, and look for evidence in ensuing talk that will confirm or reject these hypotheses. Thus, they seem to make scoring decisions in a cumulative way, rather than always orientating towards single instances. 38

3.4.2 Examiners can make a decision at almost the same time, but orient to completely different features of candidate talk. They can assign different scores for different criteria. 39

3.4.3 Pronunciation issues may influence examiner decisions in relation to other criteria.... 41

3.4.4 Occasionally, when a candidate's speech is unclear, examiners may assign a score for a criterion other than pronunciation. Yet the notes they add on the VEO app may still indicate that the observation is in relation to pronunciation............ 41

3.4.5 It is difficult to identify the cause of disfluency – is it due to lack of grammar, vocabulary, shortage of ideas or nerves? 42

3.4.6 During stimulated recall interviews, examiners may sometimes hear the same recording differently from during the rating procedure, may notice different features, and as a result may wish to change, add, or remove their scoring decisions. 43

PART 2 44

3.5 What are your feelings about the scoring experience of using the VEO IELTS tagset with the video, in comparison to the real IELTS Test examining? 44

3.6 How similar and how different was the experience of rating with the VEO tagset to the real thing? 45

3.7 Would examiners score any differently in the two test types? 46

3.8 What are the relative advantages of using the VEO tagset? 46

3.9 What are the relative disadvantages of using the VEO tagset? 47

3.10 Do examiners prefer the original IELTS rating method or the VEO tagset procedure?............ 47

3.11 Would using the VEO IELTS tagset with a video be a good way of training IELTS speaking test examiners? 48

3.12 Any suggestions for improving the VEO IELTS tagset system? 48

3.13 Do you prefer tagging and making notes on VEO as you go along, or would you prefer to do so after watching the video? 49

3.14 Answers to Research Questions 2 and 3 49

4 Conclusions 50

4.1 Answers to the three research questions 50

4.2 Discussion 51

4.3 Future research 52

References 54

Appendix A: Each examiner's scoring decisions 56

List of tables

Table 1: Number of decisions made for each criterion 21
Table 2: Number of decisions made at each section of the test 22
Table 3: Overall scores assigned by raters for each criterion 25
Table 4: Krippendorff's Alpha reliability estimate for raters based on overall scores (ordinal data).. 26
Table 5: Krippendorff's Alpha reliability estimate for raters based on decision times (nominal data) 26
Table 6: Evidence comparing examiners A and D's scoring of fluency and coherence for Lilly....... 29
Table 7: Evidence comparing examiners A and C's scoring of grammar for Lilly 30
Table 8: Examiner C's grammar scores and notes for Zoe 31
Table 9: Evidence comparing all examiners' scoring of lexical resource for Lilly 31
Table 10: Evidence comparing all examiners' scoring of lexical resource for Zoe 33
Table 11: Evidence comparing examiners A and B's scoring of lexical resource for Zoe 34
Table 12: Evidence comparing examiners A, C and D's scoring of pronunciation for Zoe 35
Table 13: Evidence comparing all examiners' scoring of pronunciation for Zoe 36
Table 14: Evidence comparing all examiners' scoring of pronunciation for Zoe 37
Table 15: Evidence of examiners B's scoring of grammar for Lilly 38
Table 16: Evidence of examiner B's scoring of fluency for Zoe 39
Table 17: Evidence comparing examiners B, C and D's scoring of different criteria for Lilly............ 40
Table 18: Evidence of examiner C's scoring of grammar for Lilly 41
Table 19: Pronunciation issues may be implicated in scoring decisions in other areas 41
Table 20: Examiner B identifies a grammatical cause of disfluency with Zoe 42
Table 21: An Examiner request for grade amendment during the stimulated recall interview 43

List of figures

Figure 1: Viewing scoring decisions on VEO 14
Figure 2: Making scoring decisions on VEO 14
Figure 3: Adding notes on VEO 15
Figure 4: Post-tagging on VEO 15
Figure 5: Research design 16
Figure 6: Comparison of all examiners' decision points and scores for both candidates................. 20
Figure 7: Interactive interface for scoring episodes 22
Figure 8: Data presented for each scoring episode 23
Figure 9: All decisions made by all examiners for all criteria for candidate Lilly 24
Figure 10: All decisions made by all examiners for all criteria for candidate Zoe 25
Figure 11: Decision points for fluency and coherence for Lilly and Zoe 28
Figure 12: Decision points for grammar for Lilly 29
Figure 13: Decision points for grammar for Zoe 30
Figure 14: Decision points for lexical resource for Lilly 31
Figure 15: Decision points for lexical resource for Zoe 32
Figure 16: Decision points for pronunciation for Lilly 34
Figure 17: Decision points for pronunciation for Zoe 34
Figure 18: A's decision points for Lilly demonstrating relationship between time and score............. 56
Figure 19: B's decision points for Lilly demonstrating relationship between time and score 56
Figure 20: C's decision points for Lilly demonstrating relationship between time and score 57
Figure 21: D's decision points for Lilly demonstrating relationship between time and score 57
Figure 22: A's decision points for Zoe demonstrating relationship between time and score............. 57
Figure 23: B's decision points for Zoe demonstrating relationship between time and score 58
Figure 24: C's decision points for Zoe demonstrating relationship between time and score 58
Figure 25: D's decision points for Zoe demonstrating relationship between time and scores 58

1 Introduction

The aim of the research is to investigate which specific features of candidate talk IELTS Speaking Test (IST) examiners orient to when taking scoring decisions. This is enabled by the development of an IST scoring scheme for a video app. This creates a recording of when exactly in the test the examiners have noticed specific features of candidate talk and taken specific decisions on scoring as a result; the decision-making is further explored in stimulated recall interviews. We also research whether the use of the scoring scheme and customised app potentially adds any value to IST examiner development and the IST rating process.

The rationale for the study is as follows. A previous study (Seedhouse et al., 2014) provided a number of findings on how features of candidate discourse relate to scores allocated to candidates by examiners. However, the problem is that we are lacking clear evidence of which features of candidate discourse examiners **actually** use as evidence in practice for deciding on scores in the four bands whilst conducting the IST. The limitation of research so far has been that there has been no way to confirm that the features hypothesised as being noticeable to examiners were, in fact, the ones which were actually noticed by the examiners who scored those tests. The research gap is therefore: which specific features of candidate talk do examiners orient to and utilise for decision-making on grades during the IST?

In this project, we employ a customised scoring scheme and app to document four examiners' decisions on scoring the four bands when viewing the test as a video recording. The technology is explained in Section 1.4. The app with customised IELTS tagset enabled recording of which specific features of candidate talk examiners oriented to, and utilised for decision-making on grades. Examiners themselves recorded this information by marking their scores onto the video at the precise point when they notice a specific feature of candidate talk. The graphical representation of their decisions alongside their comments during stimulated recall protocols formed the evidence-base of our arguments. The examiners recorded **what** they noticed **when**, and **how** they evaluated the talk. We developed two graphical means of representing their 'noticing trajectories', as well as their degrees of convergence and divergence. Stimulated recall and focus group interviews enabled further investigation of **why**, namely the possible reasons for convergence and divergence.

1.1 Literature review

This study builds on existing research in two areas (Seedhouse & Nakatsuhara, 2018). Firstly, research which has been done specifically on the IST, as well as on oral proficiency interviews (interview tests) in general. Secondly, it builds on existing research into the specific issue of how features of candidate discourse relate to scores allocated to candidates. The first of these areas is historically represented by a broad range of research methodologies, approaches, and interests, from investigations into test-taker characteristics to cognitive, context, scoring and criterion-related validity (Taylor, 2011). However, the interest in the relationship between candidate speaking features and their scores did not came to the fore till the late 1980s, as researchers turned to the question of the authenticity of interview tests (Weir et al, 2013). This interest was initiated, in part, by van Lier's (1989) now seminal call to investigate the interaction which takes place in the interview test. Fifteen years ago, Lazaraton (2002: 161) noted that there had been very little published work on the empirical relationship between candidate speech output and assigned ratings. Since then, a number of studies have added to our understanding in this regard.

It is now widely recognised that it is important to know how candidate talk is related to scores for a number of reasons. Test developers may use discourse analysis of candidate data as an empirical basis to develop rating scales (e.g. Fulcher 1996, 2003; Nakatsuhara, 2014). Similarly, evidence of the relationship between candidate talk and grading criteria can provide valuable input for validation processes (e.g. Brown, 2006a; Brown, Iwashita & McNamara, 2005; Iwashita & Vasquez, 2015; Tavakoli, Nakatsuhara & Hunter, 2017). An empirical description of the architecture of a speaking test can be useful in verifying validity and in determining whether the interaction is as it was envisaged to be or not (e.g. Galaczi, 2013; Gan 2010).

Brown (2006a) developed analytic categories for three out of the four rating categories employed in the IST and undertook quantitative analysis of 20 ISTs in relation to these analytic categories. Her overall finding (2006a: 71) was that 'while all the measures relating to a scale contribute to the assessment on that scale, no single measure drives the rating'. Instead, a range of characteristics contribute to the overall impression of the candidate's proficiency. Brown's study identified a number of discourse features in advance and then searched for these in the ISTs in her sample, using a quantitative approach. Brown, Iwashita and McNamara's (2005) large-scale, two-phase study was interesting in that they firstly identified conceptual categories that expert judges (university-based ESL/oral communication skills specialists) attended when evaluating learners' performances on TOEFL Speaking tasks. The second study then examined the extent to which descriptions and evaluations provided in the judges' verbal reports were actually reflected in candidates' discourse. They analysed 200 candidate performances with 30 measures related to linguistic resources, phonology, fluency and content, to explore how these measures could differentiate candidates across five levels. Echoing Brown (2006a), they concluded that 'each of the selected variables, while the differences across level were real – that is, not attributable to chance – any one taken in isolation was not particularly strong in determining the overall score for the speaker' (Brown et al., 2005: 83).

More recently, incorporating insights from SLA studies, Tavakoli, Nakatsuhara and Hunter (2017) focused on fluency characteristics of candidates' performance on the Aptis Speaking Test. Their analysis included three speed measures, 12 breakdown measures and four repair measures on one's fluency to offer a comprehensive understanding on the fluency construct measured by the Aptis Speaking Test. Their analysis on 128 speech samples identified criterial fluency features that differentiate candidates across four different levels (A2–C1), suggesting ways in which the current Aptis rating scale descriptors can be modified. In addition to these *a posteriori* validation studies for existing rating scales, quantification of candidate's output language across different levels was also proven to be useful at the *a priori* validation stage, that is, when developing rating scales (e.g. Fulcher, 1996; Nakatsuhara, 2014). When speaking rating scales were developed for the Test of English for Academic Purposes (TEAP) in Japan, pilot performance data were firstly assessed using draft rating scales informed by the CEFR and expert judgements. These performances were then scrutinised by various linguistic and discoursal measures that correspond to assessment areas described in the draft rating scales. The analyses were therefore used to finalise the rating scales based on empirical evidence (Nakatsuhara, 2014).

Other researchers have applied qualitative methodologies to interview test talk. Lazaraton (2002) presents a CA approach to the validation of interview tests, suggesting that qualitative methods may illuminate the process of assessment, rather than just its outcomes. Lazaraton's (1998) study of the previous version of the IST examined 20 tests and compared the relationship between candidate talk and ratings. Findings were that: there are fewer instances of repair at higher levels; higher scoring candidates use a broader range of expressions to speculate; grammatical errors are more common in lower bands, and complex structures in higher bands; appropriate responses are more common in higher bands, as is conversational discourse.

Focusing on interactional competence displayed by candidates on paired speaking tasks of the Cambridge English General examinations, Galaczi (2013) carried out CA on 41 paired performances of B1–C2 levels to identify interactional features that played a role in distinguishing between these proficiency levels. The most salient features that showed differences across these levels included: Topic development organisation (specifically degree of topic development, topic extensions of 'own' vs. 'other' topics); Listener support moves (backchannelling, confirmation of comprehension); and Turn-taking management (in a no-gap-no-overlap manner, following an overlap/latch, following a gap/pause). She also quantified some interactional features, and demonstrated both qualitatively and quantitatively how learners at different levels display their interactional competence when they co-construct paired test discourse.

Whilst Galaczi (2013) looked at interactional competence observed in dialogic test tasks, the focus of Iwashita and Vasquez's (2015) and Iwashita, May and Moor's (2017) studies was the features of discourse competence observed in monologic test tasks. For example, Iwashita and Vasquez (2015) analysed both qualitatively and quantitatively 58 speech samples on IST Part 2 (individual long turn) across IST Band 5.0–7.0 in terms of cohesive and coherence devices and lexical richness. Their mixed-methods research revealed the features of discourse that clearly distinguished higher and lower level candidates (e.g. use of a wider range of conjunctions and more accurate use of referential expressions by higher level candidates), while other features did not seem to differentiate candidates across levels (e.g. ellipsis and substitution, use of reference).

Seedhouse and Harris's CA (2011) study of the IST found that the characteristics of high-scoring and low-scoring tests in relation to topic are as follows. Candidates at the higher end of the scoring scale tend to have more instances of extended turns in which topic is developed in parts 1 and 3. There is some evidence that very weak candidates produce short turns with lengthy pauses in part 2. Confirming Nakatsuhara's (2012) findings, the study suggests a correlation between test score and occurrence of trouble and repair: in interviews with high test scores, fewer examples of interactional trouble requiring repair are observable. This confirms Lazaraton's (1998) finding in relation to the previous version of the IST. Candidates gain high scores by engaging with the topic, by expanding beyond minimal information and by providing multiple examples, which enable the examiner to develop the topic further. Candidates with low scores sometimes struggle to construct an argument and a coherent answer. High-scoring candidates develop the topic coherently, using markers to connect clauses. Candidates with a high score may develop topic using lexical items which are less common and which portray them as having a higher level of education and social status. Candidates who achieved a very high score typically developed topics that constructed the identity of an intellectual and a (future) high-achiever on the international stage. Candidates with low scores, by contrast, developed topics in a way that portrayed them as somebody with modest and often localised aspirations. Examiners may take several features of monologic topic development into account in part 2.

1.2 The current study

Nakatsuhara et al. (2017) investigate the ratings process using a mixed methods approach and demonstrate that the contribution of multimodal aspects of interaction to the process is significant. They recommend raising examiner awareness of the use of visual information and that video mode should be employed in any possible future double-rating scheme for IST. The technological innovation of the VEO app with IELTS tagset reported in the current study now provides the opportunity to record what examiners actually notice.

Stimulated recall or verbal report methodology has been employed in a number of studies (Brown et al., 2005; Brown, 2006b; May, 2011; Nakatsuhara et al., 2017) as a means of understanding the rating process, and this would be suitable to work in tandem with the observational evidence recorded by the VEO app. In a mixed-methods study, Seedhouse et al. (2014) also identified a number of features of high-scoring and low-scoring candidate talk which examiners may notice when making scoring decisions. However, they concluded that the limitation of research so far has been that there has been no way to confirm that the features hypothesised as being noticeable to examiners were, in fact, the ones which were actually noticed by the examiners who scored those tests. This research gap provided the motivation for the current study, which combines use of video, the VEO app and stimulated recall together with graphical presentation.

1.3 Background information on the VEO app

The aim of the research is to investigate which specific features of candidate talk examiners orient to when taking scoring decisions. Before describing the research design, some technological information is necessary to explain how this area may now be researched.

The project uses the existing VEO app for iPad, information about which can be found on http://www.veo-group.com/ – free demos can be downloaded. This has been developed by a spin-off company at Newcastle University. VEO combines powerful video reflection with clear feedback data to transform collaborative professional learning. VEO's unique video tagging creates lightbulb learning moments, building effective and practical continuous improvement for students, trainers and assessors. VEO's video tagging enhances reflection through visible feedback, empowering practice-led training and assessment, improving learning at all levels from student, teacher, coach and moderator. Spun out from Newcastle University in 2015, VEO now works with 75 customer organisations including over 20 universities. VEO's customer base extends across 15 countries and five continents, and has user numbers well into five figures.

The iPad screen has a tagset, which enables users to record a lesson and enter tagging/scoring decisions whilst recording, or after recording by using the website portal. The project adapted it for this specific research focus, as well as for IST examiner training, by creating a customised tagset or scoring scheme and customised app, which we refer to as the IELTS tagset. This can be seen in Figure 2 below.

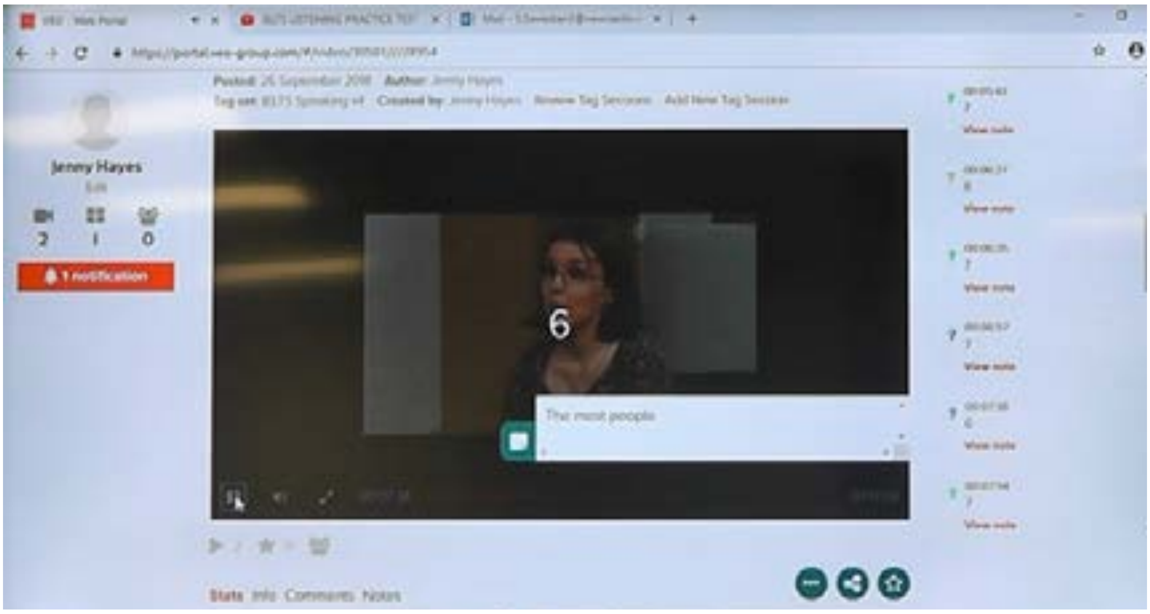
The app was adapted for this research project, as well as for IST examiner training, in the following way. The new IELTS tagset has four drop-down menus to represent each of the four IST Band Descriptor columns. Each menu features the numbers 2–9 for scoring options. If, for example, the examiner hears from the start some pronunciation problems, s/he may choose 5 on the pronunciation scale. If the candidate produces some impressive relative clauses, by contrast, the examiner may then choose 8 on the grammar scale. So this creates a recording of when exactly in the test the examiners have noticed which specific features of candidate talk and taken which specific decisions on scoring as a result. Written notes can also be added at the same time as the rating, or later. These can record why the examiners have taken these decisions, so we can know which specific features of talk the examiners were orienting to when they took their scoring decisions. The precise format of the tagset was developed by VEO staff in collaboration with the PI and Cambridge Assessment English staff at a meeting in Cambridge in order to be of maximum value for a) explication of the rating process, and b) examiner development. The technological innovation, therefore, enables research of which specific features of candidate talk examiners orient to when taking scoring decisions, as well as their reasons for doing so. The research design section below explains the procedures which were used.

As well as enabling the research element, the new scoring scheme and app also has potential for IST examiner development and the rating process, and this was a second strand which was developed in the project as research questions.

1.4 How does the IELTS tagset work with the VEO app?

Here we use screenshots to show how examiners record their grades and notes, and also how the videos appear when played back.

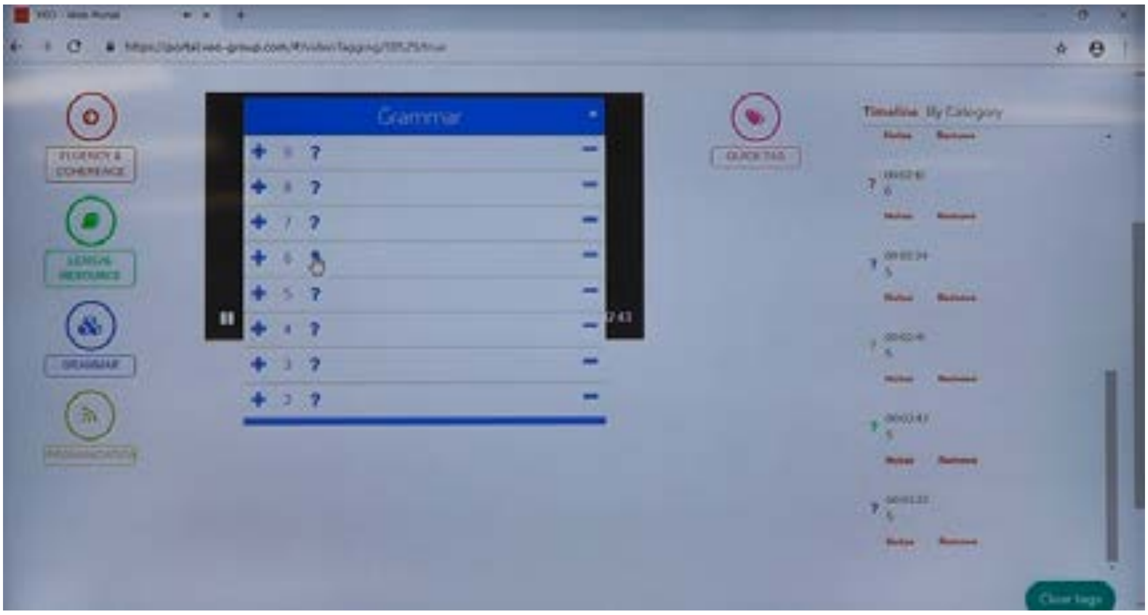
Figure 1: *Viewing scoring decisions on VEO*



In the above screenshot, we can see that at this exact point (07.38) the examiner has given a score of 6 for grammar and has written a note to show that s/he noticed the phrase 'the most people'. The right-hand column records the progress of the examiner's scoring with precise timings, with the colour indicating which of the four criteria is involved; the notes made are clickable.

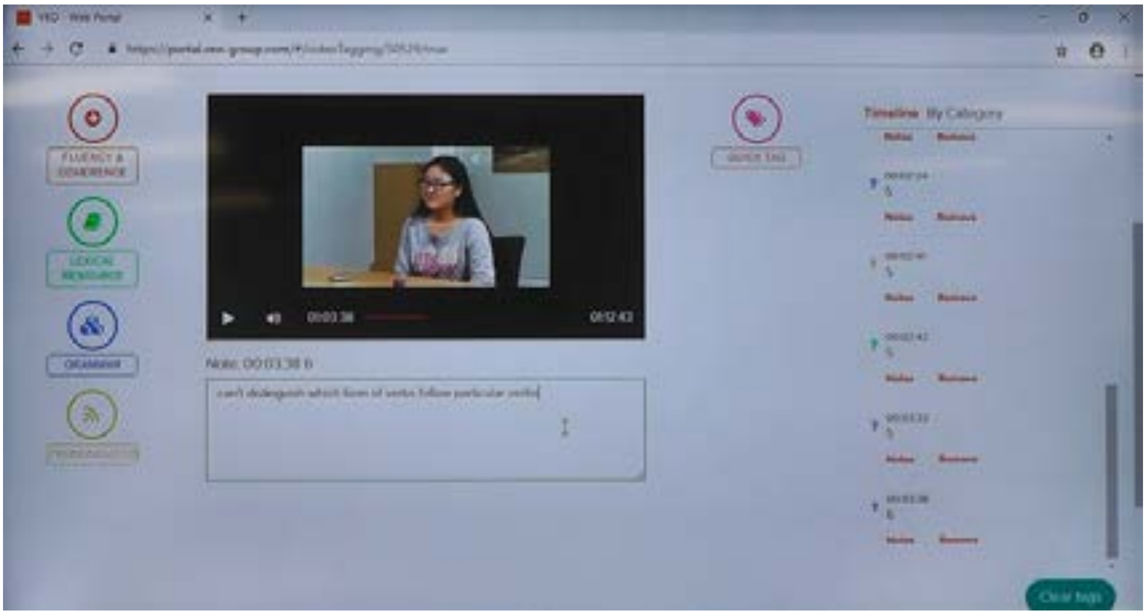
In Figure 2, we see how examiners make a scoring decision using a drop-down menu, in this case, 6 on blue for a grammar score.

Figure 2: *Making scoring decisions on VEO*



In Figure 3, we see an examiner adding notes to justify a scoring decision.

Figure 3: *Adding notes on VEO*



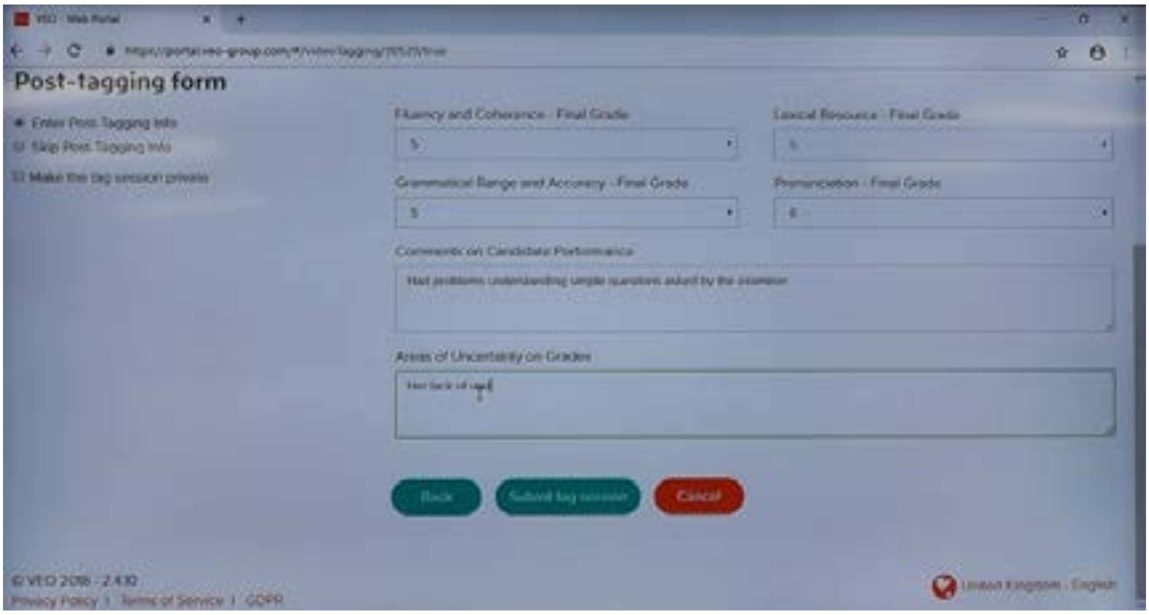
Note: 00:03:38 6

can't distinguish which form of verbs follow particular verbs

On the post-tagging form, examiners enter their final, overall scores, comment on performance and discuss areas of uncertainty.

All information entered by the four examiners was analysed and presented as data in relation to the research questions in Section 3.

Figure 4: *Post-tagging on VEO*



Post-tagging form

● Enter Post Tagging Info
○ Skip Post Tagging Info
□ Make the tag session private

Fluency and Coherence - Final Grade: 5

Lexical Resource - Final Grade: 5

Grammatical Range and Accuracy - Final Grade: 5

Pronunciation - Final Grade: 6

Comments on Candidate Performance

Had problems understanding simple questions asked by the examiner

Areas of Uncertainty on Grades

Her lack of spe

2 Research design and procedures

The research questions are:

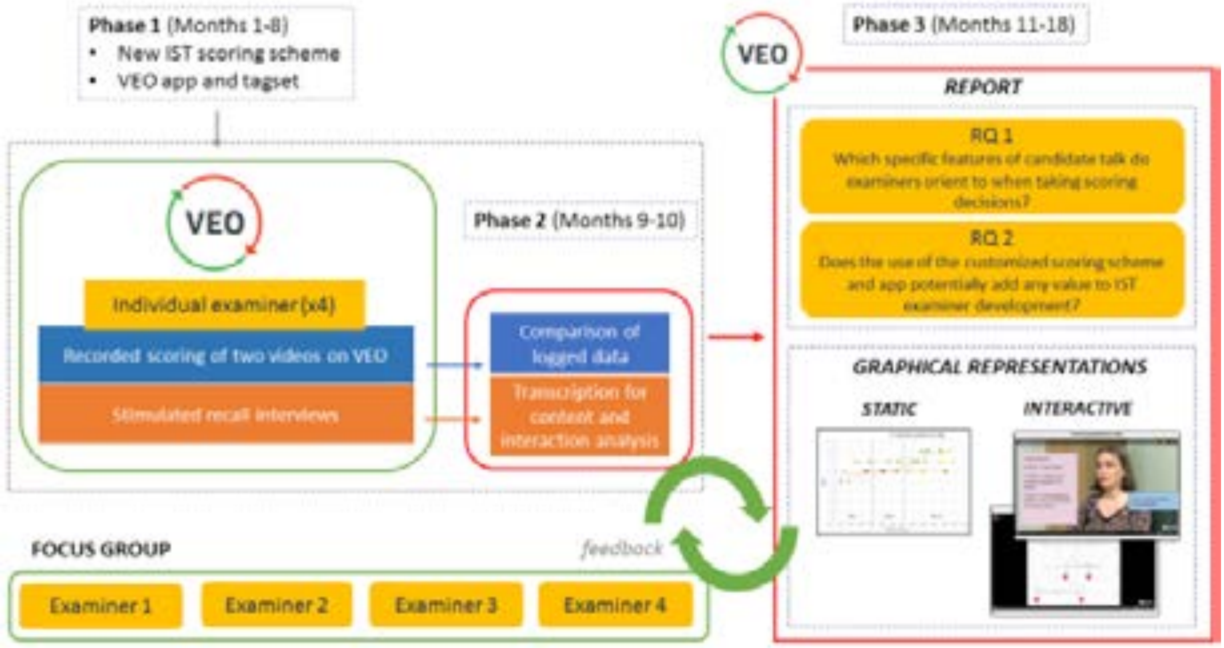
RQ1. Which specific features of candidate talk do examiners orient to when taking scoring decisions?

RQ2. Does the use of the customised scoring scheme and app potentially add any value to IST examiner development?

RQ3: Does the use of the customised scoring scheme and app potentially add any value to the IST rating process?

The overall research design was as follows. Use of the VEO app by the examiners involved the generation of both quantitative data (scores and timings) and qualitative data (notes). The data were presented for analysis in both static and interactive graphical formats. This was a mixed methods design in which the examiners' ratings and notes using the VEO app were fed back to them twice for comment as a double loop; once as a stimulated recall individual interview, and once as a focus group. The design is represented in Figure 5.

Figure 5: *Research design*


Phase 1 (Months 1-8)
• New IST scoring scheme
• VEO app and tagset
VEO
Individual examiner (x4)
Recorded scoring of two videos on VEO
Stimulated recall interviews
Phase 2 (Months 9-10)
Comparison of logged data
Transcription for content and interaction analysis
VEO
Phase 3 (Months 11-18)
REPORT
RQ 1
Which specific features of candidate talk do examiners orient to when taking scoring decisions?
RQ 2
Does the use of the customized scoring scheme and app potentially add any value to IST examiner development?
GRAPHICAL REPRESENTATIONS
STATIC
INTERACTIVE
FOCUS GROUP
feedback
Examiner 1
Examiner 2
Examiner 3
Examiner 4

The procedures followed to answer the first research question were as follows.

1. We obtained two digital videos of ISTs from Cambridge English to exemplify different score bands and trained the four examiners in how to use the VEO app with IELTS tagset for scoring.

2. Subsequently, each of the four examiners independently viewed each test using the app and scored it using the scoring tags, which logged exactly when each examiner made each scoring decision, as well as any notes made.

3. Immediately after the scoring, we conducted individual stimulated recall interviews (Gass & Mackey, 2017) with the examiners using data logged by the app showing the rating decisions taken at specific moments by the examiners when tagging. We played back their own test videos to examiners, and the logged data showed the exact point at which they took a particular scoring decision, as well as any notes made. We asked them to explain which specific features of candidate interaction they were orienting to when they made their scoring decisions. In this way, we gathered evidence of which specific features of candidate talk examiners oriented to when taking scoring decisions.

4. The stimulated recall interviews were transcribed and analysed using content analysis.

5. We devised two graphical methods (static and interactive) of presenting and analysing the logged data on examiner decisions so that similarities and differences between examiners can be depicted (see Section 3.1). This provided insight into the extent to which examiners oriented to the same, or different, features of candidate talk when rating.

6. We devised a tabular format for combining different forms of data to show different degrees of convergence and divergence between examiners. The tables combine transcripts of test interaction, scores and notes, together with relevant quotes from the stimulated recall and focus group interviews.

7. By analysing the data generated, we wrote 17 statements which encapsulated the findings (see Section 4.1).

8. Finally, we presented the findings to a focus group of the four examiners and elicited their views in relation to the research questions. We thereby verified that the statements made in the report concerning their decisions and opinions were accurate. The focus group interaction was transcribed and analysed using content analysis. A strength of the methodology is that all statements included in the research have been verified by the examiners themselves in the focus group and amendments made to any statements with which they did not fully agree.

The second and third research questions were answered by:

- stimulated recall interviews with the examiners involved in the trials
- discussions with Cambridge Assessment staff
- analysis and evaluation of all available data, which resulted in the writing of a framework and procedures for the use of the scoring scheme and customised app in IST examiner development (see Section 4.2)
- a focus group of the four examiners elicited their views in relation to the research questions.

The research project had three phases over 18 months, as follows:

Months 1–8. We developed the new IST scoring scheme for use. The precise format of the tagset and logging system was developed by the PI, RA and VEO Group staff in collaboration with Cambridge Assessment English staff in order to be of maximum value for: a) explication of the rating process; and b) examiner development. This involved a visit to Cambridge, as well as meetings by Skype. VEO Group staff then carried out the technical work to produce the tagset.

Months 9–10. Trials and data gathering as described above.

Months 11–18. We reviewed the scoring scheme and found no reason to revise it. We devised a method of presenting and analysing the logged video data on examiner decisions so that similarities and differences between examiners can be depicted in graphical and tabular formats. We analysed all data, answered the research questions and wrote the research report. When the findings were ready, they were presented to the focus group with the four examiners involved to lead a recorded discussion in relation to the research questions.

The four IELTS examiners were recruited using north-east England professional networks; two had taken part in a previous IELTS research project at the university. Their respective years of IST examining experience are: A: 12; B: 8; C: 2; D: 11 years. The examiners were given a briefing session in which we showed them how to use the VEO app with IELTS tagset for scoring and making notes. We then jointly agreed on a procedure, namely: "as you play the video, enter your score. Then stop the video and write a note of what feature(s) of candidate talk you noticed which prompted your score decision in the note space on the right side of your screen". The examiners then took a practice IELTS video home and rated it over the period of one week. The practice video was provided by Cambridge Assessment English and the purpose of the practice rating was for the examiners to become comfortable with using the new technology. Examiners all used the technology successfully to complete the live ratings, although they made some criticisms of the technology, noted in Section 3.9 below.

Each examiner then had individual sessions totalling four hours, scoring one of the two videos each time. A paper copy of the IELTS Band Descriptors was used for reference. As soon as they had finished scoring each video, the RA played back to them the scored videos and did individual stimulated recall interviews (Gass & Mackey, 2017) using data logged by the app showing the rating decisions and notes taken by them at specific moments when tagging. The RA asked them to explain which specific features of candidate interaction they were orienting to when they made scoring decisions and recorded their answers. She also asked 'naïve questions' about how they reached a decision.

The products of the research are: a) this report on the project; b) the customised scoring scheme and app for IST, which has been trialled and is ready for use in examiner development; c) a data set for the two IST videos, each of which was scored by four examiners, together with transcribed stimulated recall interviews; these may be useful for IST teacher development; and d) a graphical representation method that allowed us to investigate the scoring decisions which demonstrated the time when examiners made decisions, and the scores they assigned. This enabled identification of when examiners agreed and disagreed in their specific scores.

We then produced interactive multimodal audio-graphics which can be used by other researchers, examiners, trainee examiners, and examiner trainers to explore the data set and observe the decision-making processes of experienced examiners. This can be found in Appendix A.

3 Findings

The findings are divided into two parts. The first part answers the research question: *Which specific features of candidate talk do examiners orient to when taking scoring decisions?* To answer this question, we first present our observation for each criterion of the IST followed by findings that are not criterion- specific. The second part of the findings section answers the research questions: *Does the use of the customised scoring scheme and app potentially add any value to IST examiner development?* and *Does the use of the customised scoring scheme and app potentially add any value to the IST rating process?*

As the first section involves presentation of a novel type of data gathered using VEO, we firstly introduce some definitions.

Definitions

Criterion = one of the four IELTS bands, e.g. fluency and coherence.

Features = anything about candidate talk which has been noticed by an examiner in relation to rating.

Specific score = in relation to the VEO app, a score given at a specific point by an examiner while watching the video, in relation to specific features of candidate talk.

Cumulative score = one which builds from the start of the test up to a particular point.

Overall score = the final four band scores given by examiners at the end of the IELTS Speaking Test.

Decision point = test interaction in relation to which an examiner takes a scoring decision, but which does not meet the requirements of a scoring episode.

Scoring episode = a section of test talk in relation to which at least three out of four examiners take a scoring decision using the VEO app.

We formed the last two definitions in order to interpret the data visualised on the graphs. The decision points on the comparative graphs were varied and diverged significantly among the four examiners. In order to identify decision points that would be most suitable for an interactive display of scoring decisions, we marked episodes whereby at least three of the four examiners converged, i.e. made a decision within a time period of 10 seconds. Such episodes are termed 'scoring episodes'. These were identified, first, quantitatively, given they are within a 10 second period; and second, qualitatively, by reviewing the stimulated recall protocols to confirm whether all the decision points within the same scoring episode related to the same portion of candidate talk. Thus, while scoring episodes allowed an exploration of convergence among the examiners, the decision points that are distinctive enabled us to investigate divergence among examiners' decisions.

PART 1

Answers the research question: *Which specific features of candidate talk do examiners orient to when taking scoring decisions?*

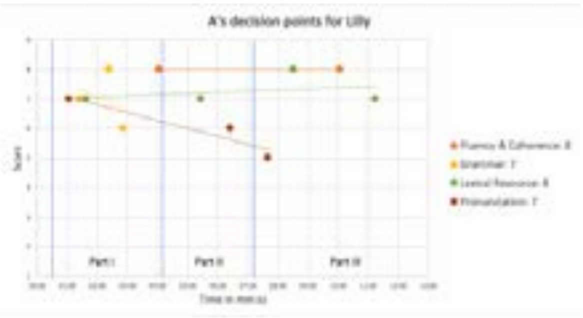
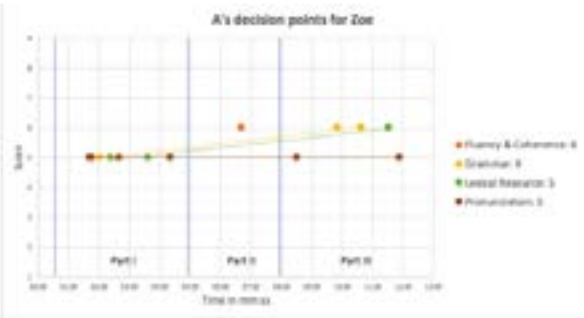
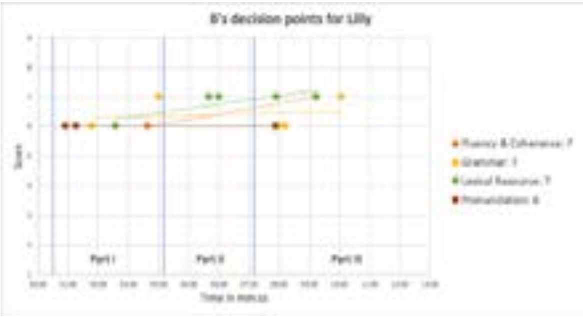
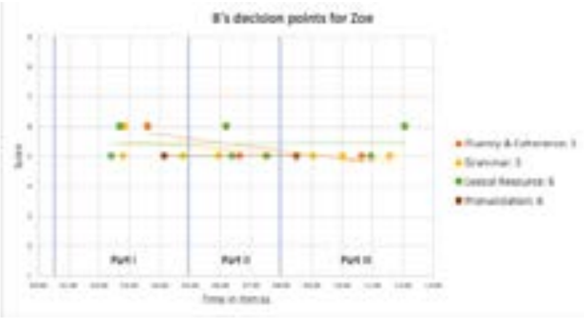
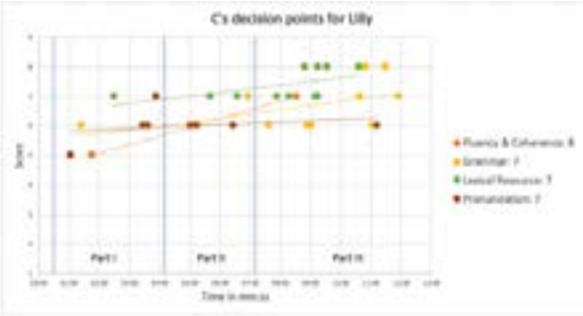
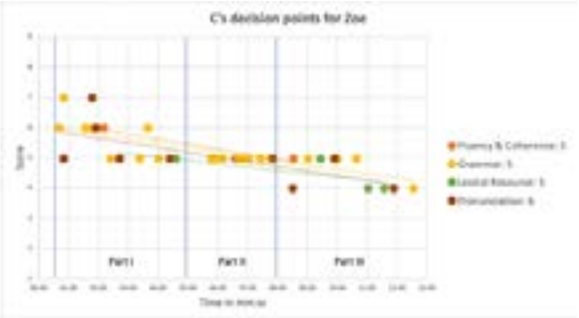
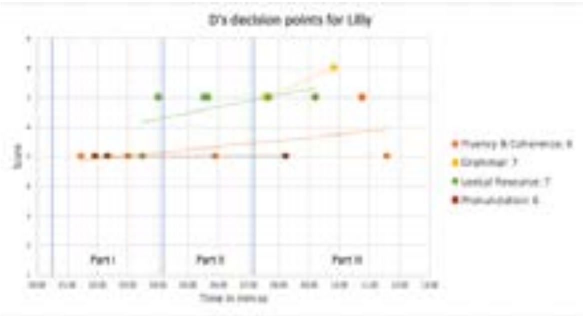
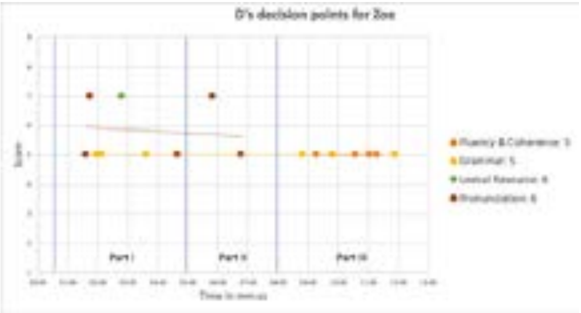
We answer this question in the following four sections. We first present static graphs of 'noticing trajectories' that show when examiners took scoring decisions, then calculate the number of decisions made for all test criteria and in each test part. We illustrate the benefits of still and interactive graphical representations of scoring decisions. The second section reports inter-rater reliability among the examiners for overall scores and decision times. The final two sections provide a total of 17 statements which encapsulate the key findings, each of which was approved by the examiner focus group. Section 3.3 investigated the salient points examiners orient to, and Section 3.4, how examiners decide when to give a score.

3.1 When do examiners make scoring decisions?

Figure 6 demonstrates each specific score given for each candidate (Lilly and Zoe) by each examiner (A, B, C, D) for each of the four criteria, as well as the times when those scores were given. The overall, final scores for each candidate are also provided in the right-hand legend for each criterion. Each figure can be found separately in Appendix A in larger format. Vertical blue lines divide the three different parts of the test. The first 30 seconds of the test involves introduction and ID check.

In general, the overall scores can be considered as the final assessment of the proficiency level of the candidate for the relevant criterion, whereas specific scores may be related rather more to noticing of specific features of candidate talk. However, as we see below, examiners also reported sometimes giving specific scores as an aggregate of 'everything up to this point'.

Figure 6: *Comparison of all examiners' decision points and scores for both candidates*

	Lilly	Zoe
A		
B		
C		
D		

According to Figure 6, examiner A has a different overall scoring pattern compared to examiners B, C and D. For Lilly, A's scores seem to stay the same or decrease as the test progresses, while B, C and D's scores indicate an upward trend over the course of the test. In the case of the weaker candidate (Zoe), examiner A's scores are again showing a different pattern from examiners B, C and D. While A's scores indicate an upward trend, B, C and D's scores indicate a downward or stable trend.

Figure 6 also suggests that examiners respond to salient points where they see evidence of certain criteria. However, as this was the first time the examiners used the technology for this purpose and were free to mark specific scores or not as they saw fit, it was to be expected that the examiners did not show internal consistency in how they managed the marking. The complexities of the decision-making process are further analysed in Section 3.4.

Table 1 shows the number of decisions made by each examiner for each candidate and each criterion. Overall, there were significantly fewer decisions for fluency and coherence (22), and decisions for this criterion seemed to be made more cumulatively compared to other criteria. This is analysed further in Section 3.3.1. The highest number of decisions were identified for grammar (59), and more than two thirds of these points were identified for the weaker candidate (Zoe). This suggests that decisions for grammar were marked in relation to specific features of candidate talk to a greater extent than decisions for other criteria. In the focus group, the examiners agreed that this was the case. This is analysed further in Section 3.3.2.

Table 1: *Number of decisions made for each criterion*

	A	B	C	D	Total
			Fluency & Coherence		
Lilly	2	2	2	5	11
Zoe	1	3	3	4	11
Total	3	5	5	9	22 (13.58%)
			Grammar		
Lilly	3	5	11	2	21
Zoe	3	9	20	6	38
Total	6	14	31	8	59 (36.42%)
			Lexical Resource		
Lilly	4	5	11	6	26
Zoe	3	7	5	1	16
Total	7	12	16	7	42 (25.92%)
			Pronunciation		
Lilly	3	3	8	3	17
Zoe	6	2	9	5	22
Total	9	5	17	8	39 (24.07%)
Total number of decisions by each examiner	**25 (15.43%)**	**36 (22.22%)**	**69 (42.59%)**	**32 (19.75%)**	**162**

While examiners A, B and D's total decisions were in the range 25–32, examiner C marked a significantly higher number of decisions (69) on the VEO app. Examiner C commented on this during the focus group as follows: "I do recall that it was, having the app made me kind of I- I wanted to record everything as- as m- I wanted to record as much as possible to help me with my decision-making, so it wasn't that I was necessarily noticing more things because I had the map- the- the app, I was just clicking I think, as much as I could in order to help me make the decision later". However, as examiners become more familiar with the VEO app, this may change as examiners develop their own unique ways of working with the app to suit their needs.

Table 2: *Number of decisions made at each section of the test*

Rater	A	B	C	D	Total
Part 1					
Lilly	6	6	7	6	25
Zoe	7	8	16	7	38
Total	13	14	23	13	63 (39.37%)
Part 2					
Lilly	2	2	7	3	14
Zoe	1	6	10	2	19
Total	3	8	17	5	33 (20.37%)
Part 3					
Lilly	4	2	18	7	31
Zoe	5	12	11	7	35
Total	9	14	29	14	66 (40.74%)
Total number of decisions in all three parts	**25 (15.43%)**	**36 (22.22%)**	**69 (42.59%)**	**32 (19.75%)**	**162**

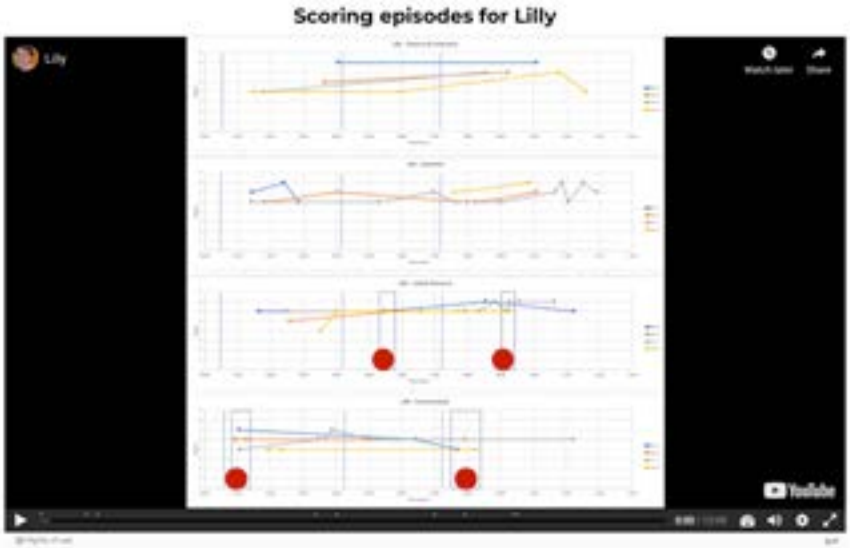
In terms of the test parts (Table 2), 80% of the decisions were made in Parts 1 and 3, and 20% were made in Part 2 of the IST.

Graphical representation of the data

The examiners' use of the VEO app and the IELTS tagset produced extremely rich data. In order to make sense of the tagged videos, we produced both still and interactive graphical representations of examiners' scoring decisions. Figure 6 exemplifies the standard 'still' graph. Figures 7, 8 and 9 exemplify interactive graphs. They demonstrate how individual decision points were mapped on scatter plots to illustrate (1) which examiner gave (2) what score (3) at what point (4) to which test criterion (5) in which part of the test. This visualisation of the decision points enabled quick comparison of scoring decisions across examiners and test criteria that demonstrated divergence and convergence. A further identification of scoring episodes indicated salient periods of candidate talk that at least three examiners oriented to for more detailed qualitative analysis.

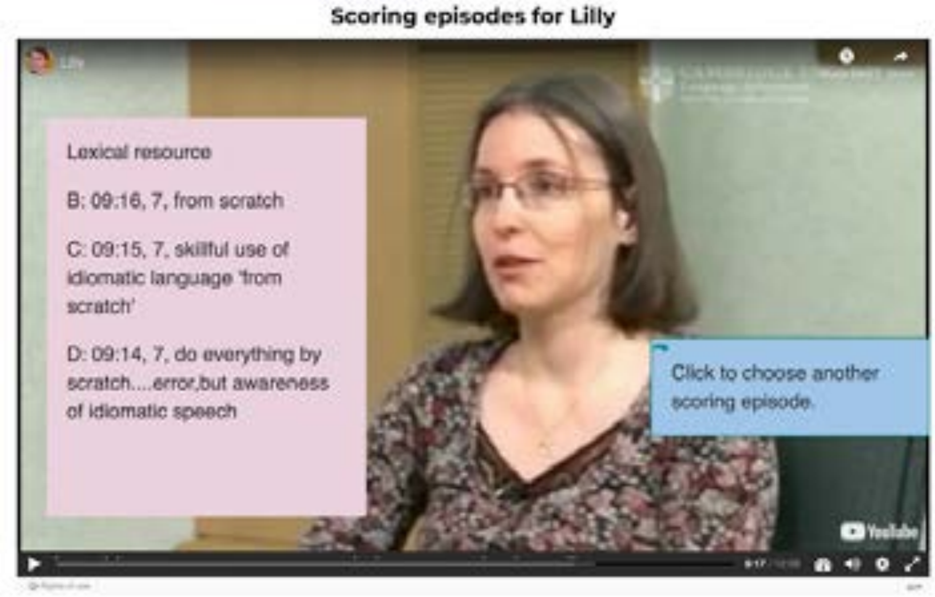
Figures 7 and 8 illustrate the interactivity enabled by this resource for one scoring episode found in Figure 9.

Figure 7: *Interactive interface for scoring episodes*



For each graph, we created an interactive interface which enables the user to explore decision points or scoring episodes. In this figure, each red round button is clickable and is linked to the specific video segment.

Figure 8: *Data presented for each scoring episode*



Scoring episodes for Lilly

Lexical resource

B: 09:16, 7, from scratch

C: 09:15, 7, skillful use of idiomatic language 'from scratch'

D: 09:14, 7, do everything by scratch....error,but awareness of idiomatic speech

Click to choose another scoring episode.

Once a scoring episode is chosen, the video plays the relevant segment of candidate talk and displays examiners' scores, the timings of these scores and their comments. The candidate talk for the episode represented in this figure is as follows: "she couldn't go to shop and buy ready meal or anything she had to do everything from scratch". The interactive graphical representation enables a direct and detailed examination of which features of candidate talk the examiners actually orient to when making scoring decisions.

3.2 Are examiners consistent in their decisions on scores and decision times?

Figures 9 and 10 illustrate all examiners' decisions (scores and times) for all criteria: fluency and coherence, grammar, lexical resource, and pronunciation. Vertical lines divide each part of the test. Scoring episodes are identified with a blue rectangle. These episodes indicate sections where at least three examiners have made a decision within a 10 second period. Each criterion and related decision-making processes are explained further in Section 3.3.

Figure 9: *All decisions made by all examiners for all criteria for candidate Lilly*

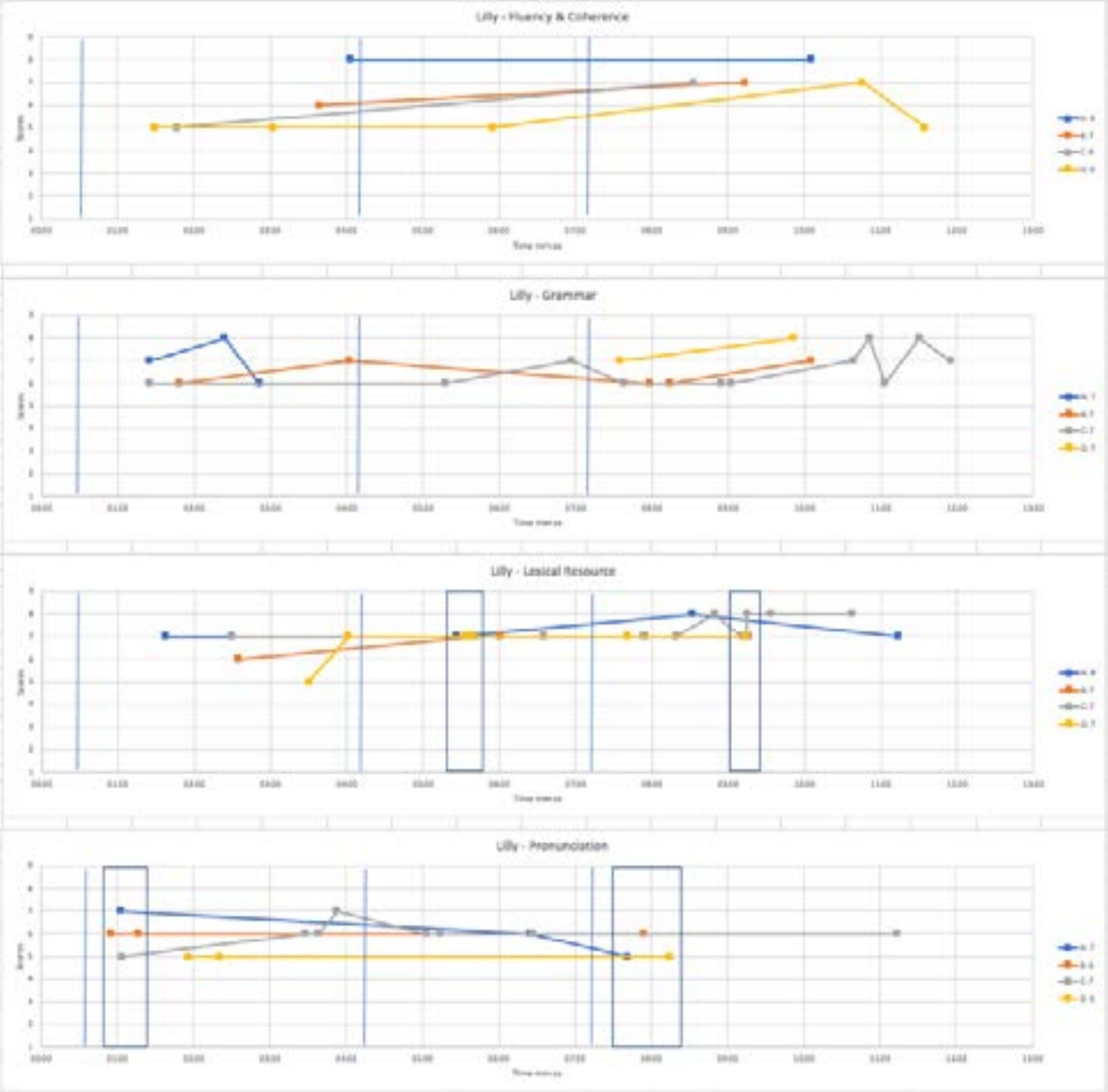
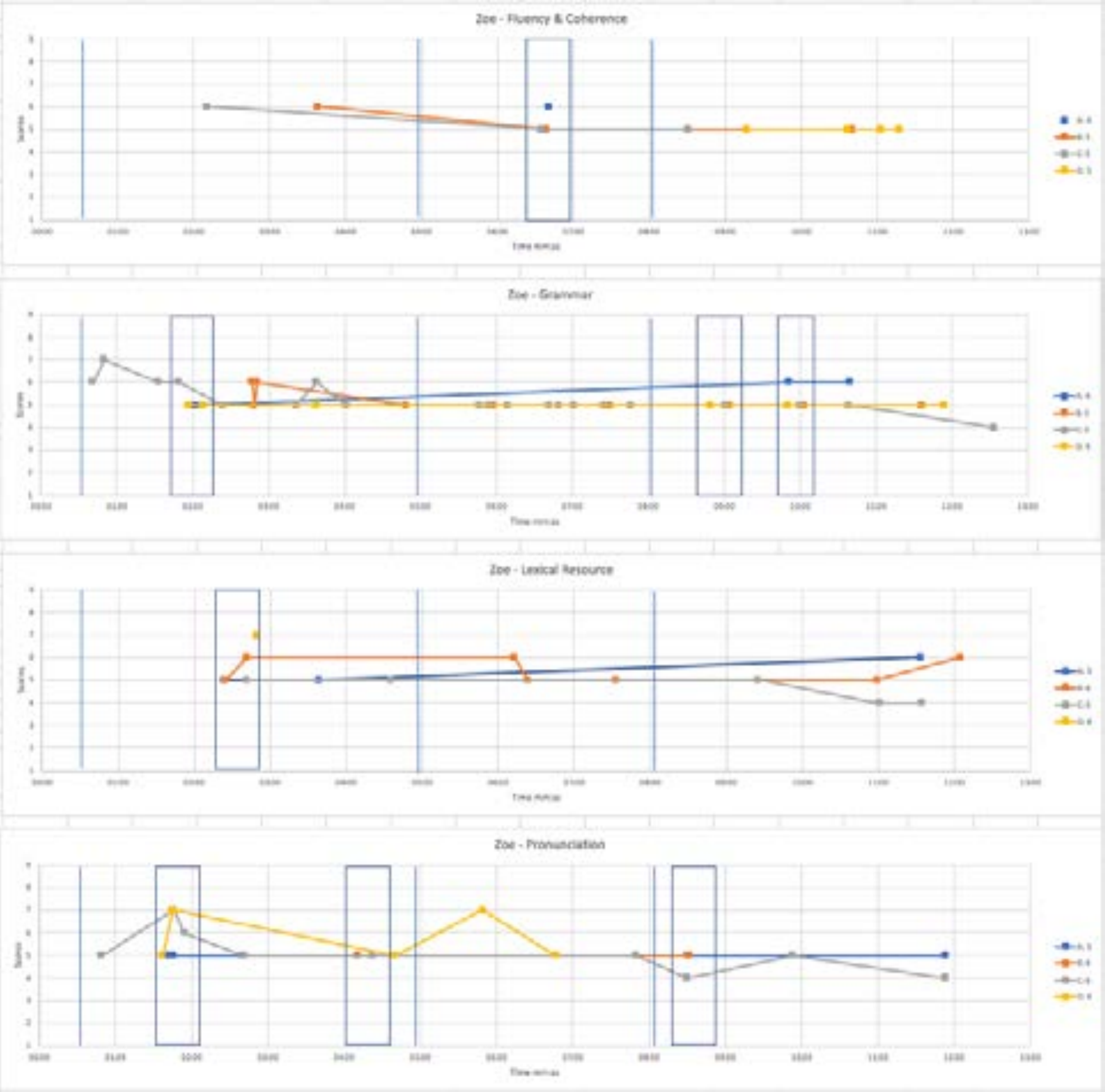


Figure 10: *All decisions made by all examiners for all criteria for candidate Zoe*



3.2.1 Inter-rater reliability for overall scores

The Krippendorff's alpha was calculated (Hayes & Krippendorff, 2007) to estimate overall inter-rater reliability scores (Table 4) between the four examiners for overall scores assigned to each criterion (Table 3).

Table 3: *Overall scores assigned by raters for each criterion*

	Rater A	Rater B	Rater C	Rater D	O (Official scores)
Zoe-fluency	6	5	5	5	5
Zoe-lexical	5	6	5	6	6
Zoe-grammar	6	5	5	5	5
Zoe-pronunciation	5	6	6	6	5
Lilly-fluency	8	7	8	6	8
Lilly-lexical	8	7	7	7	8
Lilly-grammar	7	7	7	7	8
Lilly-pronunciation	7	6	7	6	7

Table 4: *Krippendorff's Alpha reliability estimate for raters based on overall scores (ordinal data)*

Observers	Alpha	alphamin	Q	Units	Pairs
A-B-C-D	.6864	.6700	.3668	8	48
B-C-D	.7780	.7000	.1771	8	24
A-B-C-D-O	.7256	.7000	.2331	8	80
A-O	.8101	.8000	.3655	8	8
B-O	.7848	.7000	.0765	8	8
C-O	.8077	.8000	.3600	8	8
D-O	.6796	.6700	.4406	8	8

Table 4 shows that inter-rater reliability between the examiners A, B, C and D was (α = 0.68), i.e. that the four examiners were in agreement at an acceptable level. When examiner A was removed, the reliability score increased to .77, which indicated that scores assigned by examiners B, C and D were more consistent with each other.

Table 4 also demonstrates inter-rater reliability scores when official IELTS scores are taken into consideration. The scores indicate that when official scores are included as a fifth rater, all raters are in agreement (α = 0.72). Inter-rater reliability between each rater and the official scores also indicate a good level of reliability, especially for raters A, B and C (α = 0.81, α = 0.78, α = 0.80 respectively). The reliability score for examiner D is acceptable (α = 0.67), and the difference seems to be due to examiner D's scores for Lilly, which are one or two bands lower than the official scores. In the focus group, the examiners all agreed that the inter-rater reliability was 'within the parameters' and was similar to patterns they had experienced in previous examiner moderation meetings. In the focus group, we also asked the examiners: *What do you notice when you compare your individual scoring decisions with your final scores?* All examiners agreed that this seemed to them to show a pattern of consistency in their scoring practice, within their own parameters. They did not see wild movements in either direction.

The answer to the question: *Are examiners consistent in their decisions on scores?* is yes, they are, to an acceptable level. Only 1 of the 32 overall grades given by the examiners in total is more than one grade away from the official IELTS score.

3.2.2 Inter-rater reliability figures for decision times

Inter-rater reliability figures for decision times were calculated to identify whether raters noticed the same salient candidate behaviours, regardless of the criterion or score. In order to do this, a nominal value was assigned to every potential decision point as raters marked their decisions on the VEO app. In other words, a value of 1 or 0 was assigned to indicate whether a decision was or was not made by a rater at every second of the test. The initial sections for ID check that lasted for about 30 seconds were not calculated, and data from both candidate's videos for all four raters resulted in a total of 1407 potential decision points (seconds). Table 5 shows Krippendorff's alpha scores for all raters at two different intervals, a potential decision at every second, and at every 10 seconds.

Table 5: *Krippendorff's Alpha reliability estimate for raters based on decision times (nominal data)*

Observers	Alpha	alphamin	Q	Units	Pairs
A-B-C-D	.6864	.6700	.3668	8	48
(every second)	.0553	.5000	.9042	1407	8442
A-B-C-D	.7256	.7000	.2331	8	80
(every 10 seconds)	.1205	.5000	.9991	141	846

According to Table 5, Krippendorff's Alpha reliability estimate for raters based on decision times were extremely low (α = 0.05) even when decisions were coded at 10-second intervals (α = 0.12). This was not unexpected given the high number of potential decision points. Moreover, there is not necessarily a direct link between the time a rater notices a feature of candidate behaviour and the time the rater presses a tag to mark a decision. So it is theoretically possible that all raters notice the same behaviour at exactly the same time, but press a tag at a range of decision points (seconds) if at all. As we see below, raters reported sometimes noticing features, but not pressing a tag at that point. Rather, they gave specific scores later on in a cumulative way as an aggregate of 'everything up to this point'. Clearly, it is not feasible for raters to stop and record a decision every time they notice a feature of candidate talk and they must make individual decisions on this. The instructions given to the raters were simply to record decisions as they saw fit. As we saw in Table 1 above, there was great variation amongst the four examiners on how many decisions they took. If, in future studies, raters are given more specific instructions on when to record decisions, this might result in greater consistency between raters on the timing of recording of decisions. The answer to the question of whether examiners are consistent with regard to the timing of their decisions is: no, not at all, for the reasons given above.

As we will show below, there were few salient points to which at least three examiners oriented. In other words, there were few periods (scoring episodes) at which at least three examiners made decisions. However, although these points were quantitatively few, qualitative analyses of these points yield interesting findings as to how examiners make decisions, as well as insights into the rating process, as we see in Section 3.4.
In Section 3.3 below, we focus on each criterion in turn, illustrate the ways in which examiners made their scoring decisions and identify the salient points the examiners oriented to. In Section 3.4, we focus on evidence as to how examiners decide when to give a score across all criteria.

3.3 Which salient features do examiners orient to?

In this section, decision points from Figures 9 and 10 above will be explored in relation to each scoring criterion by reference to candidate speech at the time, examiners' scores and notes on VEO, as well as examiners' comments during the stimulated recall interview. The purpose of this exercise is to determine which factors might influence convergence and divergence of examiners' scores of the same candidate talk.
We examine each band in relation to Lilly and/or Zoe and provide argument statements to encapsulate the key points, all of which have been approved by the examiner focus group. The data are presented firstly as a combination of static graph and table formats. This is followed by a table format which is intended to provide a triangulated portrayal of the episode from the perspectives of multiple examiners, in order to depict degrees of convergence and divergence. The tables combine transcripts of test interaction, scores and notes, together with relevant quotes from the stimulated recall and focus group interviews.

3.3.1 Fluency and coherence

A close inspection of the decision points marked on the VEO app resulted in two arguments in relation to this criterion as explained in the following sub-sections.

3.3.1.1 Fluency & coherence scores are mostly assigned cumulatively, and there is often a weak correlation between specific and overall/final scores.

Figure 11: *Decision points for fluency and coherence for Lilly and Zoe*

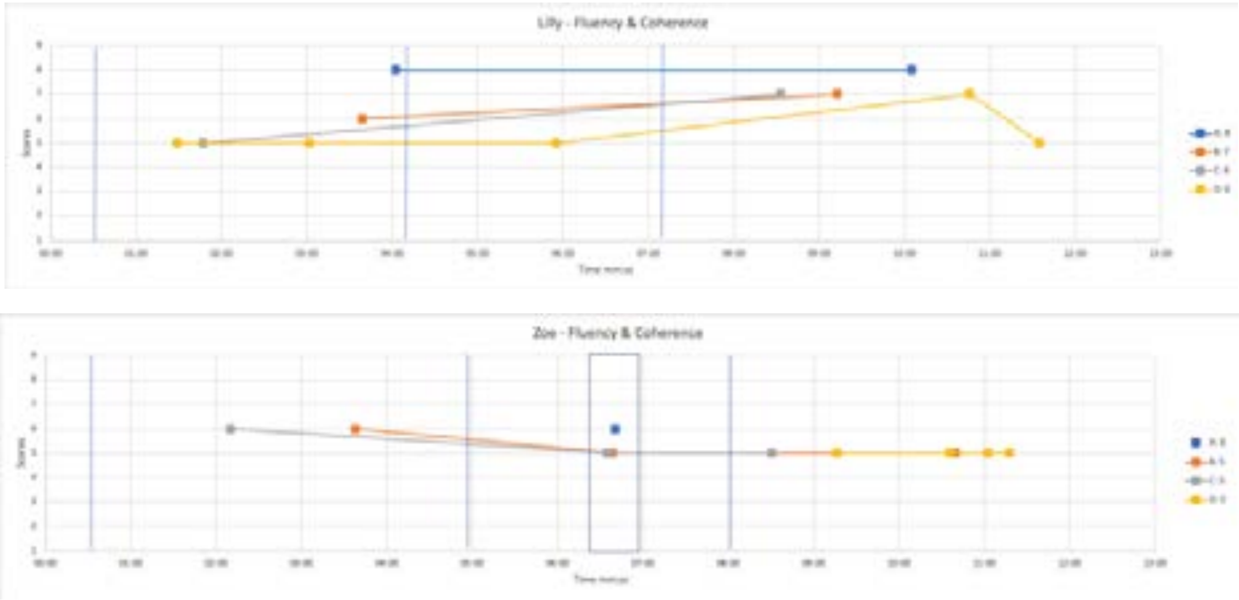


Table 1 above showed that examiners marked the lowest number of decision points (22) on the VEO app for fluency and coherence, compared with the other three criteria. In Figure 11, we further see that examiners made decisions at different points for this criterion. There is only one scoring episode in Figure 11, i.e., we do not have many salient points where at least three examiners converge on the same moment.

In the focus group, all examiners agreed that fluency and coherence scores are mostly assigned cumulatively, and the specific points marked on VEO do not always represent this criterion. There is often a weak correlation between specific and overall/final scores for fluency and coherence. Examiner B stated that "fluency is something that you cannot judge on a single err stretch".

Moreover, the specific scores do not always correspond to the overall score. For example, while examiner C only has two decision points for Lilly, which are scored at bands 5 and 7, his/her overall score is 8. As s/he explains during the interview, the fluency and coherence scores on VEO do not correspond well to this criterion. When the reasons are inspected closely in the interview, it is apparent that the first decision point might actually relate more closely to a different criterion – pronunciation or lexical resource. Reflection on the second decision point indicates that the hesitation in this instance was not related to speaking skills, but to inability to remember content.

3.3.1.2 For fluency and coherence, when examiners assign higher scores, they seem to focus on positive evidence. However, when they assign lower scores, they seem to focus on negative evidence.

Figure 11 shows that examiners A and D diverge in their overall scores for Lilly for fluency and coherence (8 and 6 respectively). While A focuses on lack of hesitation markers, and good use of discourse markers and connectives, D focuses on long pauses, and gives a lower score because fluent speech is not sustained throughout (Table 6).

Table 6: *Evidence comparing examiners A and D's scoring of fluency and coherence for Lilly*

Examiner A	Examiner D
00:04:03 Fluency & Coherence: 8	**00:03:02 Fluency & Coherence: 5**
Interview Transcript A: very fluent very little hesitation she's… it's confirmed what I thought you know that was from the first part INT: Yeah A: that- that her fluency was going to be good ermm…	VEO Note: Long pauses after initial response (searching for words?) Interview Transcript INT: and this was a five in fluency and D: Hmm… yeah erm… it's uhh… some hesitation and um the- the- the her- her turns are not very long. And it definitely seems as if she's searching for- for language. She's using these pauses to do that.
00:10:05 Fluency & Coherence: 8	**00:10:46 Fluency & Coherence: 7**
VEO Note: discourse markers and connectives Interview Transcript: A: she uses a range of discourse markers and connectives very ermm accurately. … yeah it's just- it's just very coherent and erm naturally connected speech I don't- heheh	VEO Note: Fluency definitely improves towards the end and in the interaction in part 3. Interview Transcript: D: and err you know this was just a glimpse of a seven but she doesn't really show enough of it throughout the whole test
Overall final score 8	**Overall final score 6**
Interview Transcript INT: Is there anything that you would like to change? A: No I think I would leave the fluency coherence as eight	Interview Transcript INT: So after watching the video again D: Yeah yeah INT: Umm would you change any of your scores or add anything to your comments? D: Nope. Nope. I'm sure about that one. INT: Alright. D: Umm yeah

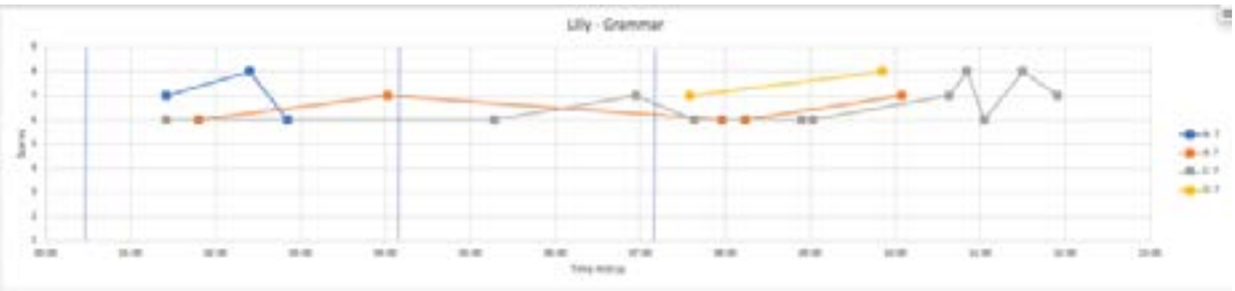
In the focus group, all examiners agreed that when examiners assign higher scores, they seem to focus on positive evidence. However, when they assign lower scores, they seem to focus on negative evidence.

3.3.2 Grammatical range and accuracy

In this section, based on the decision points assigned by the examiners on the VEO app and the resulting scoring episodes for the criterion grammatical range and accuracy, we propose two arguments.

3.3.2.1 For grammatical range and accuracy, examiners may take a decision on grammar at exactly the same time, but they may assign different scores as they focus on different features of candidate talk.

Figure 12: *Decision points for grammar for Lilly*



In Figure 12, there are no scoring episodes for Lilly's grammar scores. However, both examiners A and C take a decision on grammar at exactly the same time (01:25), but give different scores to the grammar point (7 and 6). While A focuses on accuracy in sentence structure and assigns a cumulative score that represents accuracy that has been observed 'so far', C orients to the inappropriate use of the article and assigns a specific score (Table 7). This is in line with argument 3.3.1.2 above for fluency and coherence, namely that when examiners assign higher scores, they seem to focus on positive evidence. However, when they assign lower scores, they seem to focus on negative evidence.

Time	Test interaction transcript	
01:14	E:	will you move to another town in the near future?
	L:	I don't think so but I- I'd like to- I'd like to move to Cambridge bu- eventually but-
	E:	why?
01:21	L:	because that's where I work so
01:22	E:	Hmhm
	L:	I will be right there. Because of the life.
01:25		

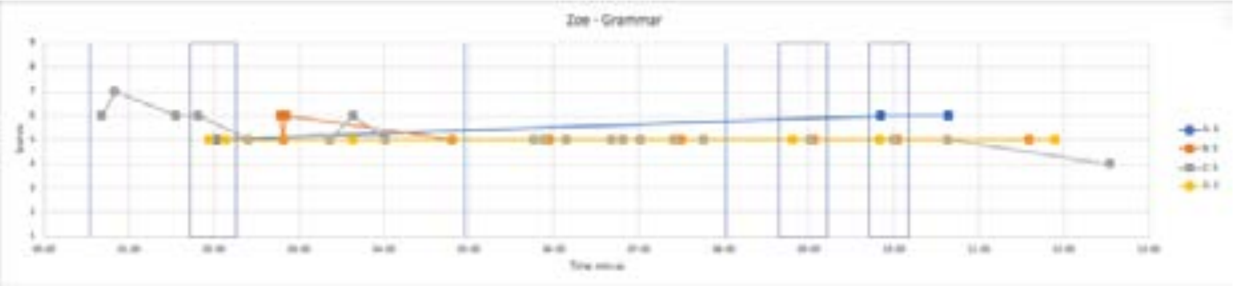
Table 7: *Evidence comparing examiners A and C's scoring of grammar for Lilly*

Examiner	A	C
Time	01:25	01:25
Score	7	6
Note on VEO	accuracy in sentence structure	Inappropriate use of article
Interview Transcript	INT: and this one was a grammar tag A: Er yes so far er I noticed that she makes- there's been very few grammatic- it's be- been high grammatical accuracy so far and sentence structure	INT: okay so this is a six for grammar C: Err it was the- the article

This is a clear example of divergence by two examiners in relation to the same criterion at the same time; one focuses on positive evidence and one on negative evidence. Both examiners agreed with this statement in the focus group.

3.3.2.2 Grammar seems to be a criterion that can be more easily tagged in relation to specific features, especially compared with fluency and coherence.

Figure 13: *Decision points for grammar for Zoe*



As we presented in Table 1, examiners marked the highest number of decision points on the VEO app for grammar (59 in total), which is an indication that examiners orient to specific features of candidate talk more often for grammar compared to other criteria, especially in comparison to fluency and coherence. In Figure 13, we observe that the grammar scores Examiner C assigns at the beginning of the test fluctuates relatively more frequently compared to other markers. This fluctuation seems to relate to a focus on specific evidence of grammar, rather than a cumulative approach (Table 8).

Table 8: *Examiner C's grammar scores and notes for Zoe*

Zoe, Grammar, Examiner C: fluctuating scores and explanations
00:00:41 Grammar 6
Note on VEO: Missing article and preposition
00:01:33 Grammar 6
Note on VEO: Limited flexibility 'do' some things
00:01:49 Grammar: 6
Note on VEO: Unable to use third person but may be a pronunciation issue
00:02:24 Grammar: 5
Note on VEO: Incorrect use of conditional may cause confusion to unsympathetic listener

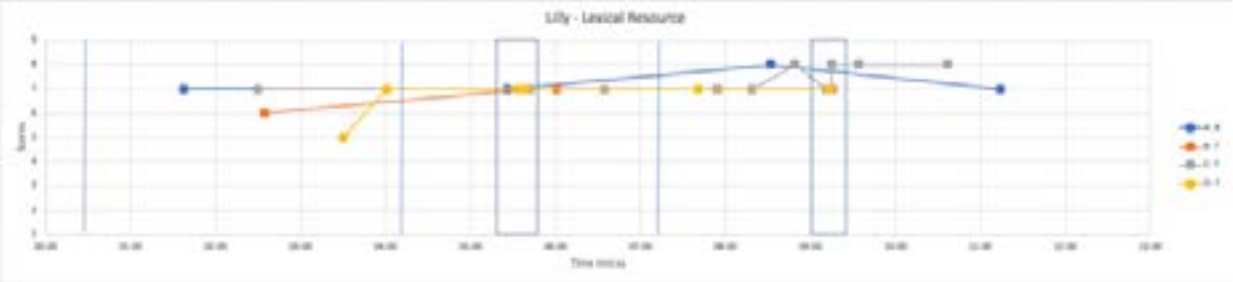
In the focus group, all examiners agreed that grammar is a criterion that can be more easily tagged in relation to specific features, especially compared with fluency and coherence.

3.3.3 Lexical resource

In this section, we present three arguments in relation to examiner behaviour as they assign scores for the criterion 'lexical resource'. While the first argument is based on the scores assigned to Lilly (Figure 14), the other two relate to scores given for Zoe (Figure 15).

3.3.3.1 Examiners tend to notice idioms and reward their use with a high mark, even if they are not delivered quite perfectly.

Figure 14: *Decision points for lexical resource for Lilly*



There are two scoring episodes in Figure 14, both of which relate to idioms ("my own boss", and "from scratch"), and receive a score of 7 from all examiners in these data. The examiners' rationales for these decisions are presented in Table 9.

Table 9: *Evidence comparing all examiners' scoring of lexical resource for Lilly*

Time	Test interaction transcript			
05:22 **05:27** **05:39** **5:42**	I do lots of uh craft I do crocheting umm (ice folding) err embroidery on cards so… anything really umm I make clay figures. Umm and it relaxes me because it's a solitary work. So… and I'm my own boss when I'm doing it haha umm…			
Examiner	**A**	**B**	**C**	**D**
Time	00:05:26	00:05:42	00:05:42	00:05:39
Score	7	7	7	7
Note on VEO	good range	my own boss, clay figures, embroidery	flexible use of expressions 'I'm my own boss when I'm doing it'	I'm my own boss

Interview Transcript Examiner A	A:	a lexical resource yeah the good- good range.
	INT:	What are the specific words that you would categorise as-
	A:	well they were the- it's- it's- it's the lexical set relating to the crafts.
Interview Transcript Examiner B	B:	alright for seven I'm- I'm looking for instances of umm collocation and idiomatic phrases
	INT:	Okay
	B:	on the one hand so 'I'm my own boss' she says there
	INT:	Hmhm
	B:	and that- that's- that's err something you wouldn't necessarily find at a six you probably you know i- i- it's akin to a seven eh there are a couple of others I think which I've typed later on
	INT:	Hmhm
	B:	Also at the same time she clearly err beyond the general vocabulary that she's got she clearly has a wider range when it comes to describing her own err hobbies like cla- making clay figures for example I th- you know that's quite a specific thing
	INT:	Yeah
	B:	embroidery
Interview Transcript Examiner C	C:	Err yeah a good use of um a- idiomatic expression "I'm my own boss"
	INT:	Yeah. That's pretty self-explanatory there.
	C:	Yeah
Interview Transcript Examiner D	D:	Again that's umm "I'm my own boss" it's- it's you know it's definitely a more sort of sophisticated usage you know idiomatic almost
	INT:	Hmhm
	D:	Yeah collocation
	INT:	Yeah

In this case there is a clear convergence amongst the four examiners in relation both to marks allocated for the same criterion and noticing of the features of candidate talk related to the marks at the same time. The candidate talk here forms a cluster of both idiomatic expressions ('I'm my own boss' is noticed by three examiners) and a specialist lexical set related to craft hobbies (noticed by two examiners). The evidence from this section is that examiners tend to notice idioms and reward their use with a high mark, even if they are not delivered quite perfectly. Examiners seem to find that some idioms help them to distinguish levels as they know what level of class will have been taught particular idioms. All examiners agreed with the statement in the focus group. Examiner B noted that this "reflects the err examiner's response to the band descriptors because seven is the first stage at which these features are noticed".

3.3.3.2 'Rehearsed' or 'memorised' chunks may not receive as high a score as other 'sophisticated' idioms which are employed in a more context-sensitive way.

An exception to argument 3.3.3.1 (i.e. idioms are noticed easily and receive a high score – mostly 7) was observed in the scoring episode for Zoe: to play a significant role in one's life (Figure 15).

Figure 15: *Decision points for lexical resource for Zoe*

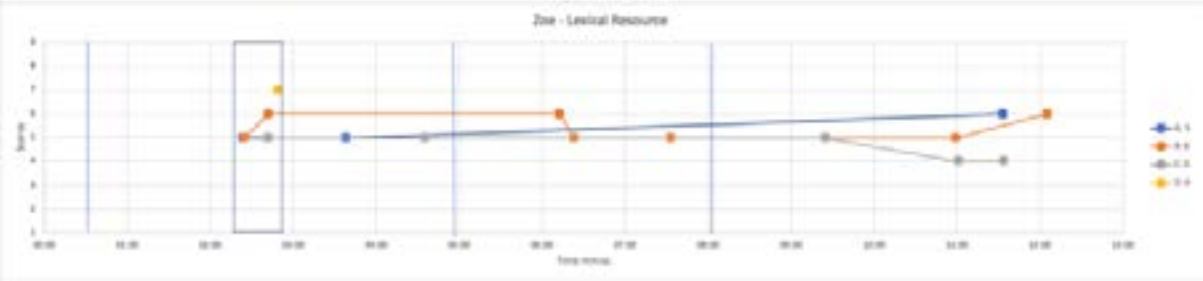


Figure 15 shows that there was only one scoring episode for Zoe that all four examiners oriented to, yet only two examiners assigned this the same score (A and C assigned 5, while B gave 6 and D gave 7) as presented in Table 10. Although both examiners B and D acknowledge that the sample sounds like a 'rehearsed' / 'memorised' chunk, they still give it a 6 or 7. Examiner A scores this evidence at 5 for wrong word stress.

Table 10: *Evidence comparing all examiners' scoring of lexical resource for Zoe*

Time	Test interaction transcript			
02:33 02:36 02:45	E: How important is happiness in life? Z: Mmm I think uh happiness is a pl- play a sin- significant role in our lives. So mm I- I- I couldn't imagine the people uh life without happiness so			
Examiner	A	B	C	D
Time	00:02:40	00:02:42	00:02:42	00:02:49
Score	5	6	5	7
Note on VEO	word stress – significant	play[s] a significant role	Significate [sic] role	I couldn't imagine.... happiness play a significant role in peoples' lives

Interview Transcript Examiner B	B: right now she's got a bit of a an idiom there INT: Hmmm B: to play a sig- play a role which is correct and the uhh adjective significant again umm again it's something that has been rehearsed perhaps err it's something that she's- she's learnt and uh brought into the- the dialogue but still it's- INT: Hmhm B: it's uh it's evidence of uh of some knowledge of uhh broa- broader knowledge of idioms so
Interview Transcript Examiner D	INT: This is a lexical seven D: Yeah there're just a couple of examples umm… `I couldn't imagine` INT: Yeah D: People dadada and happiness plays and `a significant role in people's lives` okay it's probably memorised but it's er it's- INT: Ehehe D: You know you have to mark what you hear right INT: I see D: This is quite impressive structure INT: Alright

The evidence is therefore that examiners can distinguish between idioms which appear to 'belong to' the developing topic and sequence and memorised idioms which are 'imported' to the talk in order to impress. Examiners suggest that the distinction is consequential for grading. All examiners approved this statement in the focus group. Examiner B noted that "they've rehearsed it and rehearsed it and very often you're looking at the um err this the grammatical fit into the sentence to see if they're just parroting this or or whether it actually is you know, they've they've got it linguistically umm blended into the structure of the sentence".

3.3.3.3 When the candidate is not able to use simple vocabulary, this might be noticed by the examiners, and marked with a lower score.

Our second argument based on the vocabulary decisions made for Zoe is based on another chunk ('Too high' – instead of 'too expensive') oriented to by two examiners (A and B), which receives the same scores, which are assigned at almost exactly the same time, and are given the same explanation by the examiners during the interview (Table 11).

Table 11: *Evidence comparing examiners A and B's scoring of lexical resource for Zoe*

Time	Test interaction transcript	
02:20	maybe uh you want to buy some food it's too- it is- it's too highs ermm but er er you couldn't buy it	
Examiner	**A**	**B**
Time	00:02:24	00:02:25
Score	5	5
Note on VEO	is too high – limited range for a familiar topic	too high i.e. expensive
Interview Transcripts	INT: It's a lexical point A: Yeah it's- it's like 'you want to buy some food it's too high' sh- she… it's like she's- she's got quite a restricted range for… an unfamiliar topic oh well… er it's… she's yeah she- her range is- is quite limited I think	B: Yeah 'you want to buy some food it's too high' INT: hmhm B: and er would've made sense if she said the prices are too high or INT: Hmm B: something like that but it was- it was to- err yeah it was just kind of a in- inappropriate use of the adjective there yeah

This evidence indicates that when the candidate is not able to use simple vocabulary, it might be marked with a lower score. This statement was agreed by the focus group.

3.3.4 Pronunciation

For pronunciation, we present five arguments in relation to the decision points depicted in Figures 16 and 17.

Figure 16: *Decision points for pronunciation for Lilly*

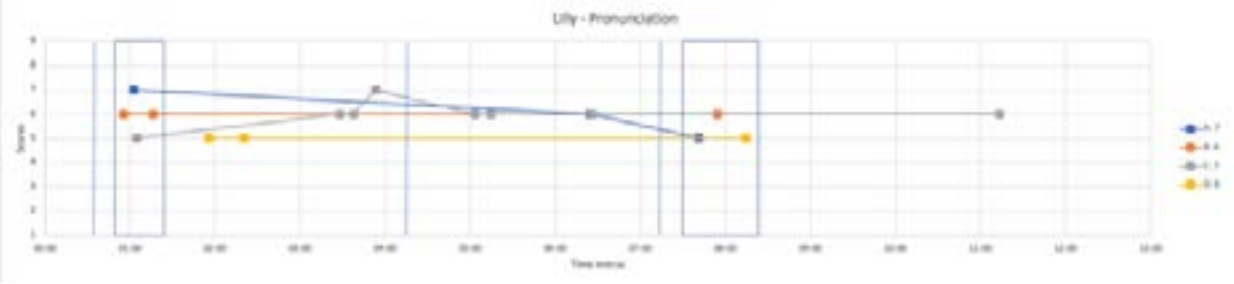
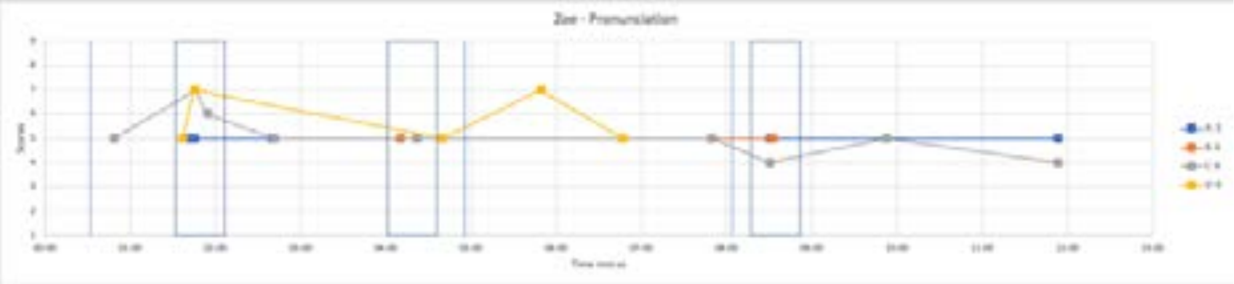


Figure 17: *Decision points for pronunciation for Zoe*



3.3.4.1 Specific scores may relate to specific problems noticed for pronunciation, whereas cumulative scores may be higher.

In Figures 16 and 17 above, we observe that examiners tend to assign specific scores for pronunciation which can be lower than the overall score. For example in Figure 17, B and C's scores are almost always 5 or under, but their overall score is 6. Similarly, in Figure 16, while examiner D marks three decisions at band 5, the overall score is 6. This may be because they give scores when they notice specific items, which are often problems. By contrast, the overall score is more cumulative. The examiners' focus group confirmed that this is how the examiners themselves explain this finding.

3.3.4.2 When examiners assign higher scores for pronunciation, they tend to focus on positive aspects (i.e. what the candidate can do); and when they assign lower scores, they focus on negative aspects (i.e. what the candidate cannot do). This provides additional support for argument 3.3.1.2.

This argument is evidenced in the first scoring episode identified in Figure 17 where three examiners (A, C and D) orient to a stretch of candidate talk, yet assign differing scores (5, 7 and 7 respectively). Examiner explanations during the stimulated recall interviews are as follows (Table 12).

Table 12: *Evidence comparing examiners A, C and D's scoring of pronunciation for Zoe*

Time	Test interaction transcript		
01:28	E: Let's talk about happiness, what makes you feel happy?		
01:33	Z: Mm I think err do some err things uh make me happy is uhmm… uh for example I'm a big fan of uh Mr Bean		
	E: Hmhm		
01:45	Z: So I really like to uh see a British er series		
01:49	E: Hmm		
	Z: So I think uh it's make me happy		
Examiner	**A**	**C**	**D**
Time	00:01:46	00:01:46	00:01:45
Score	5	7	7
Note on VEO	unintelligible words	good use of stress on 'really'	Good word stress....I REALLY
Interview Transcripts	INT: Again a pronunciation tag. Still? A: Yeah INT: Ah okay so A: (unclear) INT: oh I missed the first one A: Yeah there's some strain INT: And this is also due to the unin- A: To- to pron yeah yeah INT: Okay	C: So you know that- did I say six? That last one? It was wasn't it? No seven INT: For the `really`? C: Yeah INT: That was a seven C: No then it's… six INT: would it be a six? C: Yeah INT: and why is that? C: Err… because erm a six… er a seven would have to include… fle- ma- more things INT: Hmm C: like a range of pronunciation features but she doesn't have a wide range of pronun- she doesn't have all the positive features of- of a seven	D: Yeaah an isolated example of a sort of something that would- put the pronunciation up if consistent you know word stress for emphasis and I'm `really` li- I really like. Umm… which is often a problem with candidates the intonation is just flat

In Table 12, we see that in relation to the same extract, examiner A focuses on negative aspects, namely unintelligible words with a score of 5, whereas C and D both pick out positive aspects, namely good word stress on 'really', for a pronunciation mark of 7. The examiner focus group confirmed this statement.

3.3.4.3 All examiners may notice the same salient point at exactly the same time. However, a single item may not affect the candidate's scores significantly as examiners look for sustained evidence.

The third scoring episode illustrated in Figure 17 is a clear example of an episode where all examiners notice exactly the same salient point (Table 13).

Table 13: *Evidence comparing all examiners' scoring of pronunciation for Zoe*

Time	Test interaction transcript	
08:06	E:	Let's consider first of all job types, in your home town in- in China what are the main sort of jobs that people do?
08:17	Z:	Mmm I think uh maybe a teacher
	E:	Hmhm
08:21	Z:	in a secondary school or primary school it's a popular or some mm student study
	Z:	uh (axe) (ielts)
08:30	E:	Hmhm
08:43	Z:	Hmm yeah. Because I- I'm a science student I just know about some student uh uh some subject about science

Examiner	A	B	C	D
Time	00:08:31	00:08:32	00:08:30	00:08:30
Score	5	6	4	5
Note on VEO	unintelligible word	science? pron unclear	IELTS or Axe?	

Interview Transcript Examiner A	A:	but uh but they had a i- something 'ice' or '" I couldn't make out a word there
	…	
	A:	So… er eh that puts it into the fi- yeah it doesn't bring it down to the four band it keeps it in the- she's still in the five band
	INT:	Okay.
Interview Transcript Examiner B	B:	Err I couldn't- err eheh (was it science) I don't know what she was- I couldn't work out what she was trying to say.
	…	
	B:	Yes yeah IELTS yeah er well you see I- I just um err I was wai- I was waiting for uhh further clarification but
	INT:	Hmhm
	B:	It didn't eheh it didn't come. Yeah umm this is perhaps only the second or I think it's- and I think it's only the second time that I've ac- I haven't been able to understand at all.
Interview Transcript Examiner C	C:	Oh okay uh yeah it- it's not clear that she's saying IELTS
	INT:	Hmhm. Okay.
	C:	and and because she hasn't answered the question it doesn't help the listener to understand what she's trying to say. Had the question be about exams
	INT:	Yeah
	C:	then we could've kind of guessed but
	INT:	Hmm
	C:	the- the question was about jobs and her response is about studying for IELTS so
	INT:	Hmhm
	C:	the listener has a problem making that link quickly
	INT:	Hmhm. Alright.
Interview Transcript Examiner D	D:	IELTS she means, I think
	INT:	Ahh
	D:	maybe
	…	
	D:	Or science? Dunno.
	INT:	Would this be worthy of a umm
	D:	Well indi- i- mispronunciation of individual words it's causing difficulty for the listener yeah
	INT:	Hmhm
	D:	so five in the pron yeah
	INT:	again

In this case, all raters notice the same pronunciation problem and make a pronunciation scoring decision within 2 seconds of each other. The score decision ranges from 4 to 6, which shows some diversity. In the focus group, the examiners explained this diversity as to be expected at 8 minutes into the test. At this point in the test, examiners are said to be fine-tuning their scores and likely to focus on specific aspects that may be different from each other.

In the focus group, examiners C and B stated the following:

C: I think it is more looking for sustained evidence or lack of it.

B: Yeah this is er eight minutes into the interview now and er by now you're you're

C: We're really fine tuning it there yeah

B: I think the further into the interview you go the more divergence err you could expect

3.3.4.4 For pronunciation, all examiners may assign the same scores at roughly the same time. However, examiners may notice different aspects of candidate talk and provide different descriptions of the pronunciation problem.

For this argument, we draw on the second scoring episode in Figure 17 where we observe all examiners assign a score of 5 with close proximity, but in relation to different points within the same stretch of candidate talk. While examiners C and D find problems with the pronunciation of the word 'parents', A orients to the pronunciation of the word 'dictionary', and B to the phrase 'virtue and beauty in man'.

Table 14: *Evidence comparing all examiners' scoring of pronunciation for Zoe*

Time	Test interaction transcript	
04:01	Z:	Ermm it's like rose but uh it have a special mean. Uh for example it means uh **virtue and uh beauty in man**.
04:10	E:	Hm. Have you ever wanted to change your name?
04:12	Z:	Mm actually I don't want to because my my father err was uh uhh uh see a **dictionary** and uh find this name ehe
04:15	E:	Hmm. Who generally chooses a baby's name in your country?
04:22	Z:	Mmm I think uh **presents**
	E:	Hmm
04:27	Z:	Yeah father and mother
04:31	E:	What traditions for naming babies are there in your culture?
04:3 04:40	Z:	Mmm in my country, ermm parents want to uh uhhh name baby and some- some word that means uh uhh lucky or happiness

Examiner	A	B	C	D
Time	00:04:22	00:04:10	00:04:22	00:04:40
Score	5	5	5	5
Note on VEO	some unintelligible words	dubious patch	Dictionary	parents x4
Interview Transcripts	INT: And this was also a pronunciation point A: Yeah yeah 'some dictionary' (pronounced like 'dictionory') 'did some' yeah some unintelligible words. That one dict- dictionary? Diction- … A: Yeah it's strain i- th- it causes um	INT: let's just uh listen again Z: …virtue and uh beauty in man. INT: Is it just that part? … B: (whatchu and a building man) (imitating the candidate) hahaha I mean I- I'm hahaha INT: Yeah B: ermm I'm sure wi- with more context I could've worked it out but uh you know coming into it cold there	C: and she says presents INT: ooh okay see I think umm… I think she's actually trying to say parents there C: Huh. Play it again then INT: but yeah. Because she does mumble mother and father after it but… C: She did yeah	D: Yeah so 'parents' but she's saying 'prarents' or 'presents' so yeah INT: And f- this bit? D: About four times I think she says it

In this extract, all four examiners give the same score for pronunciation due to problems with comprehension. However, each provides a different narrative as to which aspect of the candidate's talk is responsible for the problems and they identify different features in the discussion. However, in the focus group, the examiners agreed that they did not see this variation as being a problem at all; they noticed the same phenomenon but expressed their description of it differently.

In the final section of Part 1, we turn to divergence among the examiners observed across various criteria in order to explore how examiners decide when to give a score.

3.4 How do examiners decide when to give a score?

In this section, we present six observations which became salient during data analysis as we explored decisions across examiners and criteria. We firstly present the transcript of test talk with timings, then list the scoring decisions taken and the VEO notes written. We then show the transcripts of relevant stimulated recall interviews with the examiners and provide qualitative written analyses. As a final stage, we verified with all of the examiners in the focus group whether the observations were correct and we have added relevant comments from that session. In this way, we provided multiple perspectives on scoring episodes and ensured triangulation.

3.4.1 Examiners start forming hypotheses as they listen to the candidate talk, and look for evidence in ensuing talk that will confirm or reject these hypotheses. Thus, they seem to make scoring decisions in a cumulative way, rather than always orientating towards single instances.

Below we present two examples to support this argument from the same examiner (B); one for grammar (Table 15), and one for fluency and coherence (Table 16). While the former illustrates how examiners notice evidence that confirms a previously formed hypothesis, the latter shows how contradictory evidence is evaluated. In this process, examiners may orient towards certain salient features, but not necessarily make a decision (tag the video with a score) right away. This observation is closely related to observations 3.3.1.1 and 3.3.2.1.

Table 15: *Evidence of examiners B's scoring of grammar for Lilly*

Time	Test interaction transcript	
09:19	E:	And what about in the future obviously you mentioned that we have lots of help with things like technology will that make our lives more or less stressful? In the future.
	L:	Um. I don't think it's the technology I think it's the expectation that makes- makes a life more stressful. Espects- expectation from a society and people around us than the technology.
	E:	In what way? What d- how do are expectations affect stress?
	L:	Um. Because I think people are expected to achieve more.
10:10	E:	Hmm
	L:	than- than before. You know like my grandma had a big garden and she worked part time that was her life that was it.
Examiner	B	
Time	10:05	
Score	Grammar: 7	
Note on VEO	complex sentences nearly always accurate, few gr errors	
Interview Transcript	INT:	and this was a grammar seven
	B:	Yeah the umm… this yeah it's just confirming what I would have err ma- remarked on a couple of earlier occasions this is just another example of quite a- er sequence er uh of well formed complex sentences uh umm…
	INT:	Alright

The interview shows examiner B had previously formed the hypothesis that the candidate was able to produce well-formed, grammatically complex sentences and the decision at 10.05, following a further such sentence, confirms the hypothesis.

Table 16: *Evidence of examiner B's scoring of fluency for Zoe*

Time	Test interaction transcript
03:25 **03:37** **03:43**	E: Hehe. Is unhappiness always a bad thing? Z: Mmm I think it's not bad thing but uh I don't like unhappiness. So mm… so if I'm unhappy I want do something make me happy yeah so… (unclear)
Examiner **Time** **Score**	B 03:38 Fluency and coherence: 6
Note on VEO	slight hesitation =?
Interview Transcript	INT: up until this point for fluency B: Yeah up until this point yeah I- I'm you know I'm s- I'm sticking with the six up to now umm … It's cumulative really it's- … And here she's now what she's into the second umm part of the first phase of the interview … at this point in the interview … she- she's being moved onto less familiar ground INT: Hmhm B: And err now we s- we're starting to see the cracks the weaknesses here INT: Hmm … B: Err I'm- I'm holding back to s- you know to get a bigger sample there. Umm… I'm very conscious that i- a- in- at the beginning she was well assured and uh she had good control. Uh now we're starting to see a f- uhh… more hesitation and more fumbling around and uh the sentences structures starting to sort of err break up a little bit … B: So I think that err that slight hesitation might be a si- just a sign of umm looking for ideas (or) searching for an answer as opposed to searching for language … B: I'm coming down- I'm giving her the benefit of the doubt here I think it's she's just looking for ideas because if someone asks you wh- haha wh- wha- is- is unhappiness always a bad thing INT: Hmm B: you know it's quite a philosophical question you might have hehehe INT: Yeah B: You might need to think about it yourself er er in your own (unclear) in your own language so INT: Alright

As we saw in Figure 11, this is the first decision examiner B has made on Zoe's fluency and coherence. Although s/he notices potential problems with hesitation, s/he gives a cumulative fluency rating of 6 up until that point, and states explicitly that s/he will be looking for evidence to confirm/disconfirm this hypothesis from then on. Examiner B's subsequent decisions, as well as overall score, for Zoe's fluency and coherence were 5, which indicates that s/he rejected his/her initial hypothesis of a score of 6. The statement was agreed by the examiner focus group, but there was some variation of opinion as to when exactly examiners start forming hypotheses about scores; we investigate this issue further in Section 3.4.2.

3.4.2 Examiners can make a decision at almost the same time, but orient to completely different features of candidate talk. They can assign different scores for different criteria.

In Table 17 below, examiners B, C and D make a decision within 5 seconds of each other. Examiner B makes a decision for the criterion grammar (6), C for lexical resource (7), and D for pronunciation (5).

Table 17: *Evidence comparing examiners B, C and D's scoring of different criteria for Lilly*

Time	Test interaction transcript
08:02	L: Whereas there are lots of clubs and cafes for older generation
08:08	E: So what do you think are the stresses for the older people?
08:13	L: Well. I think they have much less stress than the younger people
	E: Hmm
08:19	L: because as I said the community- community caters for them so heavily
08:24	E: Hmm
	L: umm like… there is a- the- senior club and the knitting club
	E: Hmhm

Examiner	B	C	D
Time	08:14	08:19	08:15
Score	Grammar: 6	Lexical resource: 7	Pronunciation: 5
Note on VEO	e.g. [the] older generation	Flexible use of vocabulary 'the community caters for them'	I missed that completely
Interview Transcript	INT: So we've got a grammar six B: Yeah that's the omission of the definite article INT: Ah B: yeah umm she m- she said 'the younger people' but uh 'older generation' she missed out 'the' uh. If yo- yeah uh contrasting	INT: Okay so these are both a seven. And… it's due to the idiom- idiomatic use? C: Yeah yeah INT: Alright C: (And the)… collocation of caters INT: Hmhm C: community caters for people as well INT: Yeah	INT: Ermm… yeah this part was a pronunciation five D: Yeah I missed- I missed um "I make-" could we just have that again INT: Yeah D: Oh okay "I think they have much less stress than" I missed it f- initially INT: Okay D: Hmhm

The data show that three of the four examiners take a scoring decision within 5 seconds of each other and comment on features of candidate talk within this extract. However, each has selected a different band for grading, given a different grade and commented on a different feature of candidate talk within the same extract. This example of divergence demonstrates that, in some cases, examiners may orient to different features of candidate talk and reach different scoring decisions, apparently as a result of these differences in noticing.

During the focus group, the examiners related this apparent diversity to the progress and timing of this episode in the test. They found it significant that this happened 8 minutes into the test. At this stage, they would have a fairly clear hypothesis as to potential grades and would be engaged in 'fine tuning' and looking for specific features to confirm or disconfirm those hypotheses. They felt, therefore, that later in the test, more divergence is to be expected in terms of them orienting to different features of candidate talk during the same sequence of the video. Examiner B observed that "the further into the interview you go, the more divergence err you could expect".

In the later stages of the test, diversity of focus and noticing of features amongst examiners may be understood in the following way. Examiners may be sure of the grade for some criteria but not others, and different examiners may be unsure of different criteria. Therefore, it is to be expected that, in the later stages of the test, examiners will focus on different criteria and notice and grade different features of candidate talk. This implies that the concept of inter-examiner reliability for specific (not overall) scores and timing of decisions needs to be related to the stage of the IST at which a decision occurs.

3.4.3 Pronunciation issues may influence examiner decisions in relation to other criteria

Table 18 shows that there was some indication that pronunciation problems have an impact on examiner decisions in other criteria. In this example, examiner C assigns a score of 6 for grammar to Lilly as s/he wrongly assumes that the candidate says “mostest people”, as opposed to what the candidate appears to intend, namely “the most stressed people”.

Table 18: *Evidence of examiner C’s scoring of grammar for Lilly*

Time	Test interaction transcript
07:23 **07:33** **07:41**	E: where you’re living now, what do you think are the main causes of stress for people? L: In Downham Market I think the causes of stress umm the most stressed people I imagine are younger people
Examiner **Time** **Score**	C 07:38 Grammar: 6
Note on VEO	The most people
Interview Transcript	C: Hmm the mostest INT: This was a grammar point C: Yeah she s- she actually said the mostest people I think INT: Yeah C: but I d- I had already written it and started again INT: Okay C: So it was the mostest INT: Okay C: But in any case it’s wrong

In the interview, examiner C suggests that s/he has assigns a score of 6 for grammar to Lilly as s/he assumes that the candidate says “mostest people” at 7.33, and identifies this as a grammatical error. The candidate appears to intend to say “the most stressed people”, which is how the transcriber heard it. In the test audio, the candidate’s pronunciation is very unclear. Another examiner (A) also had significant difficulty in the interviews in understanding what the candidate was saying here, hearing it first as ‘mostest people’ and finally as ‘most stressed people’. Examiner A also notes "it could be because of the- the recording or it may be if in a live situation, I might have understood it”. So, we should conclude that there is some evidence that pronunciation issues may influence examiner decisions in relation to other criteria. However, it is unclear whether this happens in the face-to-face live tests, or whether examiners may mishear candidates more when listening to recordings. This might be a topic for future research.

3.4.4 Occasionally, when a candidate’s speech is unclear, examiners may assign a score for a criterion other than pronunciation. Yet the notes they add on the VEO app may still indicate that the observation is in relation to pronunciation.

Table 19: *Pronunciation issues may be implicated in scoring decisions in other areas*

Criteria	Examiner	Candidate	Time	Score	Explanation on VEO
Fluency & Coherence	C	Lilly	01:47	5	Words unclear and difficult to understand
Pronunciation	D	Lilly	01:56	5	Unclear pron of individual words and words blend together
Grammar	C	Zoe	03:22	5	Several errors of grammar and pronunciation cause loss of meaning
Lexical resource	D	Lilly	03:30	5	Individual words are very affected by L1 accent

In Table 19 above, we see that scores may be given for criteria other than pronunciation while the notes on VEO show that a pronunciation item has been noticed. Examiners C and D have assigned scores of 5 for Lilly and Zoe using four different criteria, although the notes indicate that pronunciation problems are implicated in each case. Therefore, examiners' notes on VEO can be valuable in examiner training in order to make examiner decision-making processes evident to the trainees.

3.4.5 It is difficult to identify the cause of disfluency – is it due to lack of grammar, vocabulary, shortage of ideas or nerves?

The notes on the VEO app were also useful in stimulated recall interviews. Table 20 shows how examiner B reflected on his/her explanation for a grammar decision that implicated the impact of disfluency ("almost loses coherence?"). Using this note on the app, examiner B was able to explain that s/he noticed loss of coherence, but this was due to issues in grammar.

Table 20: *Examiner B identifies a grammatical cause of disfluency with Zoe*

Time	Test interaction transcript
06:35 07.30 07:37	Z: to be a lecturer err i- the- they can uhh mm studying with uh student and uh s- maybe and uh when you have a lesson you want to teach some uh some knowledge. The- this should be a (search online) and then incur some knowledge. Maybe you didn't be- you- you didn't know before and uh uhhh and uh you know in my country student is ve- err is r- is r- is really want to uhh study mm because uh uh maybe sometimes they are t- they are in the class and then they will ask uhh teacher some mm diffi- di- di- difficult question about this subject and uh the teacher uh the lecturer umm have to- to des- uhh describe to why its uh answer and uh. Z: And then gives- and then gives them err some er suggestion or advice.
Examiner Time Score	B 07:30 Grammar: 5
Note on VEO	almost loses coherence?
Interview Transcript	INT: And umm since the note is 'almost loses coherence' I mean why did you choose to tag this as a grammar point and not a fluency and coherence point? B: Umm I thi- i- it's- it's th- it's the faults with the grammar err that are putting a strain on the listener er f- for me at- at any rate INT: Hmm B: Ummm… and… but as I say I don't know if it's a symptom or whether it's a product of err of her errr either lack of ideas or shortage of ideas or or INT: Hmhm B: her lack of umm um vocabulary there but- she seems w- ermm… w- th- im- the overall impression I'm getting is that she's fine err as we saw it right at the beginning of the interview INT: Hmm B: Er on familiar ground, a well-prepared uh well- well-rehearsed almost automatic responses she was giving there INT: Yeah B: Ermm she had good control of uhh of those phra- phrases and the grammar and sh- she's threw in quite- a one two uh yeah a good umm good uses of umm verbs and auxiliaries. But what we're seeing now is I- I- I- I'm- I'm getting a distinct impression that as- as we push her further and further INT: Hmhm B: Umm this- this control is errr of- of sentence structure is breaking down

Examiners reported in the interviews problems in identifying the cause of lack of fluency, which could be due to lack of grammar, vocabulary, shortage of ideas, or nerves. Consequently, the same problems can be noticed but rated in different bands. In this case, examiner B chooses grammatical weakness as the underlying cause at this point. The statement was agreed by the focus group.

3.4.6 During stimulated recall interviews, examiners may sometimes hear the same recording differently from during the rating procedure, may notice different features, and as a result may wish to change, add, or remove their scoring decisions.

One clear example of examiner request for change during the stimulated recall interview is presented in Table 21. Here, while examiner C had marked this decision as 5 for fluency and coherence, during the stimulated recall, s/he requested a change to 7 for lexical resource. Thus, requests were sometimes related to a change in criteria, other times a change in scores, or both.

Table 21: *An Examiner request for grade amendment during the stimulated recall interview*

Time	Test interaction transcript	
01:44	E:	Do you like making other people laugh?
01:48	L:	Yes but I'm not (really) good at it so I don't try very often
01:52		

00:01:47 Fluency & Coherence: 5

VEO Note: Words unclear and difficult to understand

Interview Transcript

C: Yeah uh it would be a seven
INT: Hmhm
C: Cuz she's- she's using lots of- of good vocabulary and erm… err less common expressions
INT: Yeah
C: I'm not very good at it erm try to make people laugh things like that
INT: Hmhm
C: but- but it's her accent that's- that always (pulling) it a bit
(interruption)
INT: Um so in that case so this was marked as a fluency band
C: Hmm
INT: but then… upon review you would mark it as a lexical resource one?
C: Yeah yeah

Overall, for both candidate videos, the four examiners requested a total of 13 decisions to be changed, 24 new decisions to be added, and 3 decisions to be removed during stimulated recall interviews. The requests for changes are sometimes mentioned in the transcripts of the interviews, such as in Table 21 above. However, these requests for decision changes may have been related to the questions asked in the stimulated recall interviews, which prompted examiners to explain why they took those decisions. In the focus group interview, examiner C noted: "I- I would say that looking at the transcripts again I- I remember feeling a little bit as if I was being pushed into possibly making- changing something…I felt that having- being being questioned about my decision… Why why are you- why are you making that decision it fe- made me feel like I really did- maybe hadn't made the right decision I began to question myself, whereas my original decision was- was the intuitive one that I made and mostly kept it". Since it is possible that the requests to change decisions may have been to some extent an artefact of the types of questions asked in the stimulated recall interviews, we have not included those requests in our analyses. In our quantitative work, we have presented only the original decisions.

In section A, the research design enabled the detailed depiction of how examiners noticed which features of candidate talk when taking scoring decisions. The graphical presentation of 'noticing trajectories' together with qualitative notes enabled a portrayal of which features of candidate talk examiners oriented to, as well as the degrees of convergence and divergence between examiners. Presenting the data to the examiners in both stimulated recall and focus group interviews enabled further exploration of the rating process and the reasons for convergence and divergence. Data analysis generated 17 encapsulating statements, which were approved by all examiners.

PART 2

This part answers the research questions: ***Does the use of the customised scoring scheme and app potentially add any value to IST examiner development?*** ***Does the use of the customised scoring scheme and app potentially add any value to the IST rating process?***

The following are the themes taken from the individual stimulated recall interviews with the four examiners. Firstly, we present the original questions which were asked and then the examiners' responses. We cite selected quotations and intersperse these with analyses and comments. In the subsequent focus group with the same examiners, we presented the draft report to them. In some cases, we have added relevant quotations from the focus group interaction, identifying these as such, or have added reports of the focus group.

3.5 What are your feelings about the scoring experience of using the VEO IELTS tagset with the video, in comparison to the real IELTS Test examining?

Examiner A did not feel comfortable using the tagset as s/he is so used to the standard IELTS procedure (with 12 years' experience) in which examiners cannot make notes: "Erm… yes I feel it's umm it's- I'm not that comfortable with it because I'm… I'm so used to I got so used to… ermm… not- not being allowed to not- not writing anything down just- just really listening and… keeping it in my head and sort of… so as you go along you're- you're thinking `well yeah this could be… uh it looks like a five` you- you c- uh- er what happened what I tend to- tend to do is umm… err form an impression quite- quite soon he- you know after… s- when I became very experienced at it and then it's- it's like a confirmation almost but then sometimes in the third part, ermm… they err they sometimes go higher than- in- in some of the- some of the categories so yeah."
In the focus group, examiner A stated that using the tagset interrupted her flow.

However, examiner A did find some aspects useful: "So I found this umm… I found it useful that I could listen again to- so because the way it goes back and then you can listen to it again So th- to th- the section that you tagged I thought that was useful yes".

Examiner A "found it a little bit clunky at first because it goes back to just before (the point)" but found "that worked well but I mean it's just getting used to it." Examiner A would like to try "just doing the tagging and then writing notes" afterwards, rather than writing notes while going along, which is perfectly possible as an alternative.

Examiner B thought "it was a useful little tool actually. It certainly it got me thinking more explicitly than umm you know than- than the- the speaking the standard uh face to face sort of interview yeah".

Examiner C said: "I think having changed one of my scores following a second viewing I think it kinda shows that really it's- it's important to have some back up so that you can review your decision or somebody can review your decision because when you're doing it you've got your fourteen minutes and you're listening and trying- and remembering everything it's umm yeah a th- it's- I think it would be good to have some back up".

Examiner D said:" err you c- you can evaluate more umm…I would say more um efficiently because you're not doing the interlocuting at the same time right? er it's (much more-) positive y- y- yeah. You can- you can really reflect more and pay more attention because you're not speaking and thinking of questions at the same time. Umm so you- you're just doing one job not two umm…yeah I think that's the main thing".

So all examiners found an aspect of using the VEO tagset beneficial, although they did not all mention the same aspect, so there was no unanimity. One examiner did not find the experience of 'writing notes as you go along' comfortable and would prefer writing them at the end.

3.6 How similar and how different was the experience of rating with the VEO tagset to the real thing?

Examiner A said: "it's very different because ermm…you're not- well you- obviously you're not the one engaging and asking the student so that's ve- that's very different er er in a way ermm…being a- a- an observer of it it's uhh…you probably notice more features ah because when you're involved in it sometimes you're thinking about the next question". A sees the VEO experience as "taking away from the actual, the authenticity of it because that's what it is at the moment it's- it's- it's an examiner and a candidate and that two way so yeah is it is- it's a different experience".

Examiner A added: "when you're actually in with the candidate y- there's a lot going on, you're watching the time. You're thinking about, umm, you're listening to their responses particularly in th- the third part which is not scripted and so you have to ask follow up questions, so you're thinking of y- y- you're reacting to them, so you're- you're trying to do a lot you know assessing and trying this hahah you know. It's quite demanding. Umm, whereas this is like oh I'll just listen and yeah so."

Examiner B said: "well the big difference of course is that umm in the IELTS exam you're both uhh th- the examiner is both uhh interlocutor leading the- leading the ummm err dis- er read the script i- but- but still leading the conversation. As well as having to err pass judgement if you like at the s- at the same time, and that's always umm been a- ah- err quite taxing aspect of err er of IELTS speaking examining compared to other types of speaking exams that er that you do where- where you're not always necessarily the uhh uh you- you may have someone in the room with you whose actually conducting the- the exam. It's to some extent, it's umm, of course it frees you up to, uhh, to reflect on your- on your judg- err on your judgements there and it might not always be a good thing to have too much ti- too much time to to reflect on err on your decisions because I think you might tend to… he- hesitate more over a final judgement ermm, er, or sometimes you know your- your gut your gut feeling is the best or it comes out best".

Examiner C said: "I mean in a normal IELTS test- test you're- you're always worried that- that you've remembered everything enough to- to give an accurate mark. Well- well I am personally you know that i- you are doing it fairly. But here you ca- you can get I- I found I was getting a little bit kind of drawn into…looking for errors and…and- and looking in too much detail at- at all the things she was saying. Ermm but (unclear) but it is a- a solid thing under your belt to be able to- to justify the- the final grade that you've come up with. It's different because, umm, you're focusing on specific instances but this can be a positive thing and a negative thing at the same time".

Examiner D said: "Yeah it's- uhh it's totally not similar is it it's erm er this (VEO) is a- it's like a passive evaluation. Whereas the exam is… it's tiring umm because…but it's more- much more interactive. Umm…for…you are part of it that…ehh ub- but we're talking about evaluating right? Well…I think being a part of it can be brought into the conversation because when you're doing the exam, you're- you are a part of it and you're evaluating at the same time Mm…yeah you're- yeah so maybe this (VEO) is less- less subjective. Uhh mm watchi- it's more objective watching the video. Because when you're doing the exam our job is to facilitate as much as possible candidate speech right. So unconsciously you are…asking questions or using non- nonverbal communication to elicit that but if you're watching a video, y- you don't care and you just- you just eheh ehe you're a- you're a voyeur alright you're just grading them from that perspective".

All examiners agree that rating videos with VEO is a very different experience to the normal testing experience. All examiners agreed how tiring/demanding the normal test is because of the dual role of interlocutor and examiner. A common theme is that the VEO tagset marking frees the examiner up to focus solely on the candidate performance. However, all examiners agreed that the possible disadvantage of using VEO is having the leisure to focus too much on the detail of the candidate's talk and too much time to reflect on it, leading to 'overthinking'. In the focus group, all examiners agreed that the VEO tagset would be significantly more useful for examiner training and certification than for marking of the IST by experienced examiners.

3.7 Would examiners score any differently in the two test types?

Examiner A said: "in terms of the- of the scoring I think that…I don't think I would've scored them any differently had I done it live."

Examiner D noted that "watching on the video may have made us umm mark the grammar more harshly than the official version because it could be that the- the interpersonal element of an exam is missing. So it has a hidden effect the- the rapport that you might have with the student".

Examiners B and C did not express opinions on this point. In the focus group, however, all examiners agreed that they would expect to score in the same fundamental way in both test types, and they expected that the overall scores would come out the same.

3.8 What are the relative advantages of using the VEO tagset?

Examiner A notes: "I found it useful that I could listen again to- so because the way it goes back and then you can listen to it again. So th- to th- the section that you tagged, I thought that was useful yes".

Examiner A adds: "you've already got the examiner in there so this would only work for somebody who is maybe training to become an IELTS examiner." It could work if there were an interlocutor asking the questions and the examiner watching the video. "Possibly a higher level of accuracy because so yeah sometimes when you're doing the real thing, you know you may miss things obviously and you don't really- you can't really listen to them again there's no time".

Examiner B thought "it was a useful little tool actually. It certainly it got me thinking more explicitly than umm you know than- than the- the speaking the standard uh face to face sort of interview yeah".

Examiner B added: "one of the err useful er aspects of this the of the tagging is that- is that you can do it at the sa- at er in- not in exactly in real time but i- if you like as as as you go along".

Examiner C notes: "if you had two examiners, then it would be umm it would be far more objective".

Examiner D says: "er it's (much more-) positive y- y- yeah, you can- you can really reflect more and pay more attention because you're not speaking and thinking of questions at the same time, umm, so you- you're just doing one job not two".
Examiner D adds: "if they got people asking them a question in a room I'm watching from outside, many exams are done like that. Even Cambridge exams are done like that, you've got an interlocutor and you've got the other ca- you've got your colleague in the corner doing that right. So what's the difference? I mean the- it's a- it's a tried and tested system and uhh I think it would be a g- a more yeah more effective, more objective."

Three examiners agreed that there would be advantages to having two people involved in the test: an interlocutor and an examiner rating the video using the VEO tagset. One examiner noted that there would be both advantages and disadvantages.

3.9 What are the relative disadvantages of using the VEO tagset?

Examiner A said s/he: "found it a little bit clunky at first because it goes back to just before (the point)" and "did not feel comfortable using the tagset".

Examiner B said: "Apart from the having too much time to (tag it) and question your decisions not immediately, no".

Examiner C said: "Ermm…that actual- the- the actual being there in the- in the exam room is very different to to doing it on a screen so, ermm…I don't know why but I had a feeling that I could ha- had I been there in the room with her, I might've been able to understand the- th- the- the pronunciation issues that were caused by her speed. I- I d- I don't know why it's a bit of a kind of just a feeling that I that I got, er, perhaps there's more sympathy and empathy going on between in a- in a real rea- err interaction with people".

Examiner C added: "but I think you get…you can- you can- by- by making all the comments all the time, you can get a little bit bogged down in the details of it".

Examiner D did not specify disadvantages.

Three out of four examiners named disadvantages, but these were different disadvantages, so there was no unanimity. However, in the focus group, all examiners agreed that the possible disadvantage of using VEO is having the leisure to focus too much on the detail of the candidate's talk and too much time to reflect on that.

3.10 Do examiners prefer the original IELTS rating method or the VEO tagset procedure?

Examiner A answered: "Erm…yes I feel it's umm it's- I'm not that comfortable with it [VEO]".

Examiner B answered: "I would like to umm…I would like to have more experience with the tagging I think before I made up my mind on that".

Examiner C said: "The original method can be quite a heavy load, ie, you have to make quick, (almost, because of the time) irreversible decisions. This method does allow you to go back and check but it would still represent a lot more time for the examiner. I'm a little undecided, but think I prefer the original as, at the end of the day, it seems we do make the right decisions".

Examiner D said: "I'll- I…listen I, umm and I think it's got a future. And umm…for- for- I don't know it's not my job to say it's got a future for IELTS but I think it's err you know some er other exams…use similar techniques don't they but umm um yeah I- I- for me, as an ex- speaking examiner, I would rather do it in the comfort of my own home watching a video than sit and do eight nine spea- err speaking tests back to back".

There are differences between the examiners on this question, with two preferring the original method, one preferring VEO and one undecided.

3.11 Would using the VEO IELTS tagset with a video be a good way of training IELTS speaking test examiners?

According to **Examiner A**: “Yeah definitely”. Examiner A points out there’s an existing “training process that you do uh…you watch some videos tests and um and you rate them”. S/he thinks that VEO “would add to the process...With this you can quite easily stop it and- and umm and look at it again yeah so yeah. I think that would work really well...Examiners are standardised every two years I think it is and uh…we could do- do it- do it for that as well, so you don’t have to go into a centre. Doing it online would be an option with this”.

Examiner B said: “Yes definitely. During re-certifications, we- we use videos like that but we tend to watch them all the way through and then discuss them afterwards or go through it. Whereas I think particularly for those who are doing it for the first time, who haven’t done it before, yeah to- to break it down like this with examples yeah would be- would be very useful but uh…when uhh examiner trainers give back reports umm… for example they- they will- caught instances from a particular interview that they have tagged they’ve highlighted as being er characteristic of a particular band and you get feedback from that note saying (unclear) er candidate says this this and and this. And- and this is just, a- a more an- er-, a good way of making this more explicit too, yeah to the trainees so yeah. I could see it could be very useful, yeah, I wish we’d had this five years ago, hahah”.

Examiner C said: “Yeah definitely...For training examiners it would be very good”.

Examiner D said: “Yeah but it is already. Because when we do recertification, we- we- we watch videos and- and then grade them you’ve got every two- uh every two years you- you- you have a day. You look at some speaking tests together and say 'what do you think it was', 'what do you think it was' and then uhh 'okay this is what it was'. You’ve got six recordings to listen to you’ve got to grade them. So…if it was a video much better. Definitely all- all those nonverbal sort of umm…f- features in it and the clarity umm… so definitely deheh yeah oh uh…because yeah absolutely videos like this clear videos umm…definitely”.

There is complete agreement amongst examiners that the VEO IELTS tagset with a video would be a good way of training IELTS speaking test examiners. Three examiners mention that the process could be used during re-certification, as well as during initial training. They stressed that this option would save a great deal of money and time if the examiners could rate the videos online at home. They suggested that the same video could be rated and commented by all examiners separately. The examiners could then watch the standardised master video which has been rated and commented on by the lead examiner.

3.12 Any suggestions for improving the VEO IELTS tagset system?

Examiner A suggests instead of having just the score number, also having some descriptors with them too. “Instead of just pronunciation fluency, erm, and coherence, you could actually uhh maybe use some of the you know like `uses a range of connectives` particular features of- of the- of the descriptors. I think that might be useful”.

Examiner B said: “No I can’t- I can’t think of any, it’s a fairly simple thing you’re given you’re given the band descriptors, er, you’re given the four er four types of tag”.

Examiner C said: “Just the buffering really”, which refers to delays in viewing the video due to the internet connection. This was not the fault of the app, but of the Wi-Fi connection. S/he also suggested a clearer tag to show which criterion or feature the note relates to.

Examiner D said: “When you click on the tag, I think the video should stop straight away, it’s a bit of a (bind) to having to. If that could be integrated, that’s the only thing I would say. Okay so instead of pausing when you click on the note but pausing when you t- click on the tag, so that you can, instead of two clicks, one click, but that’s a small thing yeah but you do cause when y- when- when you click on it, it runs on a little bit”.

Three examiners each suggest different improvements (which will be passed on to VEO Group), so there is no agreement on points needing improvement.

3.13 Do you prefer tagging and making notes on VEO as you go along, or would you prefer to do so after watching the video?

Examiner A said: “Doing the tagging and then writing notes afterwards [would] be a better option for [me]”.

Examiner B said: “Most in- intuitively it [tagging and noting as you go along] was the, you know, the the best way to do about it. Umm, I think if you waited till the end of the video then went back and made your notes, you would’ve forgotten what, uh eheh wha- a- as, you saw, eheh er, earlier”.

Examiner C said: “[Tagging and noting as you go along] seemed the most logical way to do it”.

Examiner D said: “[Tagging and noting as you go along], I see something and I want to get it down in word form to be able to- to- to explain it precisely and uh recall recall the recall what I saw”.

Three out of four examiners preferred the option of tagging and writing notes as they went along. However, one examiner did not like this way of working (see quote above Section 3.13) and would prefer writing notes afterwards. Both options are perfectly feasible using VEO and it would be possible to give IELTS examiners in training a choice of either option.

3.14 Answers to Research Questions 2 and 3

The answer to the research question as to whether the use of the customised scoring scheme and app potentially adds any value to IST examiner development is a very clear one. There is complete agreement amongst examiners that the VEO IELTS tagset with a video would be a good way of training IELTS Speaking Test examiners. Three examiners mentioned that the process could be used during re-certification as well as during initial training. They stressed that this option would save a great deal of money and travel time if the examiners could rate the videos online at home.

The answer to the research question as to whether the use of the customised scoring scheme and app potentially adds any value to the IST rating process is, by contrast, not at all clear. The examiners saw several disadvantages, as well as advantages, to its use. Asked whether they preferred the normal rating method or using VEO, two preferred the original method, one preferred VEO and one was undecided. All said that the VEO rating experience was very different.

In the focus group, all examiners agreed that the VEO tagset would be significantly more useful for examiner training and certification than for marking of the IST by experienced examiners. Therefore, it is recommended that the VEO tagset should be used for IST examiner training and re-certification. Moreover, generation of still and interactive visualisations of experienced examiners' rating the same candidate using VEO may potentially add further value to IST examiner development. However, there is no convincing evidence that there would be any benefit for it to replace the current system for marking of the IST by experienced examiners.

The study has shown there is variability across raters and it could be argued that this makes the VEO app less useful for training purposes as it is not possible to say which of the many features raters should focus on. However, it would be possible for the lead examiner to produce a master rating with notes which identifies the features which are most relevant to the scores, for training purposes.

4 Conclusions

4.1 Answers to the three research questions

In this section we provide summary answers to the three research questions:

Which specific features of candidate talk do examiners orient to when taking scoring decisions?

A detailed answer to this question was provided by the 'noticing trajectory' graphs in Section 3 combined with the examiner notes and stimulated recall interviews. Data analysis generated the following statements, which were approved by all examiners:

1. Examiners are consistent in their decisions on scores, to an acceptable level.
2. When examiners assign higher scores, they seem to focus on positive evidence. However, when they assign lower scores, they seem to focus on negative evidence.
3. Fluency and coherence scores are mostly assigned cumulatively, and the specific points marked on VEO do not always represent this criterion. There is often a weak correlation between specific and overall/final scores.
4. Grammar seems to be a criterion that can be more easily tagged in relation to specific features, especially compared with fluency and coherence.
5. Examiners tend to notice idioms and reward their use with a high mark, even if they are not delivered quite perfectly.
6. 'Rehearsed' or 'memorised' chunks may not receive as high a score as other 'sophisticated' idioms which are employed in a more context-sensitive way.
7. When the candidate is not able to use simple vocabulary, this might be noticed by the examiners, and marked with a lower score.
8. The same scores may be assigned by all examiners for the same criterion (pronunciation) at roughly the same time. However, examiners may notice different aspects of candidate talk and provide different descriptions of the pronunciation problem.
9. Specific scores may relate to specific problems noticed for pronunciation, whereas cumulative scores may be higher.
10. All examiners may notice the same salient point at exactly the same time. However, a single item may not affect the candidate's scores significantly as examiners look for sustained evidence.
11. Examiners can make a decision at almost the same time, but orient to completely different features of candidate talk. They can assign different scores for different criteria.

12. Examiners start forming hypotheses as they listen to the candidate talk, and look for evidence in ensuing talk that will confirm or reject these hypotheses.
13. Examiners seem to make scoring decisions in a cumulative way, rather than always orientating towards single instances.
14. Pronunciation issues may influence examiner decisions in relation to other criteria
15. Occasionally, when a candidate's speech is unclear, examiners may not always assign a score for pronunciation, but a score might be assigned to a different criterion.
16. It is difficult to identify the cause of disfluency, whether it is due to lack of grammar, vocabulary, shortage of ideas or nerves.
17. During stimulated recall interviews, examiners may hear the same recording differently from during the rating procedure, may notice different features and may wish to change their scores.

Does the use of the customised scoring scheme and app potentially add any value to IST examiner development?
Yes, this is very clearly the case, and this was agreed unequivocally by all examiners, who added that it would be particularly suitable for the re-certification process. Graphical representations of scoring decisions based on the data generated via the VEO app with IELTS tagset also have potential benefits for examiner development.

Does the use of the customised scoring scheme and app potentially add any value to the IST rating process?
No, there is no convincing evidence that this was the case, nor that it should replace the current system for marking of the IST by experienced examiners.

4.2 Discussion

It is clear from the data that there are instances of convergence during the tests when the examiners notice the same candidate talk features at the same time and give the same score. In this study, use of idioms proved to be an easily noticeable and scoreable feature, for example. Conversely, there are instances of divergence during the tests when the examiners notice different candidate talk features at the same time and allocate different scores in different criteria. As reported in Section 3.4.2, during the focus group the examiners related this apparent divergence to the progress and timing of this episode in the test; later in the test, more divergence is to be expected in terms of them orienting to different features of candidate talk during the same sequence of the video. It is also clear that the examiners do not take scoring decisions in the same way or with the same frequency for all four criteria, with differences being most pronounced between grammar and fluency/coherence rating processes. We therefore cannot expect the ratings process to always develop in exactly the same way for all examiners throughout the test. The actual trajectory of the ratings process may depend on: the nature of the developing interaction; the stage of the test; the criterion being employed; which features the examiners noticed first and which initial hypotheses they formed. We should also note (see statements 9 and 10) that asking examiners to focus on specific features and mark specific scores as they go along does not necessarily influence the cumulative scores they award. Examiners are well able to distinguish between the two processes and their respective purposes.

The study makes more explicit the practical reasoning processes which IST examiners go through from listening to candidate talk to allocating grades. It provides a framework and procedures for investigating which features of candidate talk examiners notice when, and how this noticing relates to scoring. Inter-rater reliability has generally been conceived as a quantitative verification procedure. However, the systems developed here enable the study of inter-rater reliability as a process in which we investigate how and why examiners notice different features of candidate talk and allocate different scores.

The illumination of the complex processes involved in rating speaking performance may also prove useful for test validation and test design.

In terms of limitations, the study was limited to data from four examiners scoring two videoed tests each and may also be understood as a 'proof of concept' study. It demonstrates that the VEO app with IST tagset can indeed be used for scoring ISTs. This process generates rich, useful data and has considerable potential for supporting examiner development. Although the sample size is small, the mixed methods research design means that a very large amount of data was generated, only about half of which has been included in the report. In principle, there are few technical limits to any possible future upscaling of the basic research design. The app can make a video available to 1,000 different examiners anywhere in the world, who can then use the IELTS tagset to score and make notes on the performance. However, whilst the 'noticing trajectory' graphs are readable with data from four examiners, they would become unreadable with larger numbers of examiners and would require a different solution for portrayal.

Another limitation was that examiners did not record all of their noticing. Examiners did not find it at all practical to stop and record a decision every time they noticed every feature of candidate talk. Examiners experienced the VEO tagging and recording as an additional cognitive load which changed their regular rating experience. What examiners noted was clearly not a complete record of all of the cognitive processes which they went through. Examiners tended to mark cumulative scores when they noticed a pattern several times and were confident of a judgement. In general terms, it is likely that the precise task and instructions given to observers will influence the extent to which they record everything they notice or not using VEO. In this project, the overriding priority for the examiners was to provide candidate scores, a task in which they were already experienced.

4.3 Future research

The research design provides a framework for future research into grading processes for speaking tests. The examiners themselves pointed out the different cognitive processes involved in live rating and rating via a video app, and these differences would be interesting to explore in the future. An unexpected finding to this research study, which came from the examiners themselves, was that the VEO/IELTS tagset innovation would be ideally suited to examiner re-certification work. We, therefore, recommend a feasibility study to examine the practicalities and possible benefits of this.

What might a framework and procedures look like for the use of the scoring scheme and customised VEO app in IST examiner development and re-certification?

1. The VEO app with IELTS tagset would be employed. This was adapted for this research project as well as for IST examiner training in the way described in Section 2. The new IELTS tagset has four drop-down menus to represent each of the four IST Band Descriptor columns. Each menu features the numbers 2–9 for scoring options.
2. No purchase of iPad would be necessary, as everything could be run from the VEO portal website. So examiners/trainee examiners anywhere in the world would be able to take part with only a PC and internet connection.
3. Cambridge Assessment English would have to take out licences with VEO group to cover the number of examiners/trainee examiners taking part in the feasibility study.
4. Cambridge Assessment English would choose a number of videos of ISTs which are suitable for the re-certification process.

5. These videos would also need to be graded and notes written by a chief examiner to create a master grading as a benchmark, against which feedback to examiners can be provided.
6. The examiners/trainee examiners taking part need some basic training in using the VEO app and IELTS tagset. This can be done using videos on the VEO website or by making a short customised training video.
7. Cambridge Assessment English load the videos onto the VEO portal website and invite the examiners/trainee examiners taking part to access them.
8. The examiners/trainee examiners watch and grade the videos and write notes.
9. The graded videos can be immediately accessed by Cambridge Assessment English.
10. The master grading can then be made available to the examiners/trainee examiners taking part.
11. Cambridge Assessment English can then provide feedback to individual examiners and the whole cohort on their performance.
12. The exercise would also generate a great deal of data on examiner grading and moderation processes, which could be used in a number of ways.

The broader implications and possibilities are as follows. The technological innovations (VEO app with customised tagset and associated graphs) provides a significant tool for examiner training and moderation in language testing. It also provides a significant opportunity in terms of the analysis and evaluation of spoken discourse in general. This is because the same video can be played to an unlimited number of users anywhere in the world, who each record what they notice and when in relation to features of the talk, which specific features of talk they orient to, and how they evaluate them, where applicable. These data are immediately available to the researcher for quantitative and qualitative analysis. Interactive graphical representations of scoring decisions can be generated, which preserve interactional complexity whilst enabling researchers, trainees, and examiners to drill down into detail.

Given that VEO tagsets are extremely flexible and easy to produce, the technological innovation has enormous potential for research into spoken interaction and intersubjectivity across methodological boundaries.

References

Brown, A. (2006a). Candidate discourse in the revised IELTS Speaking Test, *IELTS Research Reports Vol 6*, pp 71–89. IELTS Australia: Canberra and British Council: London.

Brown, A. (2006b). An examination of the rating process in the revised IELTS Speaking Test, *IELTS Research Reports Vol 6*, pp 41–69. IELTS Australia: Canberra and British Council: London.

Brown, A, Iwashita, N and McNamara, T. (2005). An examination of rater orientations and test-taker performance on English for Academic Purposes speaking tasks, *ETS Research Report Series*, RR-05-05, TOEFL-MS-29.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction, *Language Testing 13* (2), pp 208–238.

Fulcher, G. (2003). *Testing Second Language Speaking*. Harlow: Pearson Education Limited.

Galaczi, E. (2013). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests?, *Applied Linguistics 35* (5), pp 553–574.

Gan, Z. (2010). *Interaction in group oral assessment: A case study of higher- and lower-scoring students*, *Language Testing 27*(4), pp 585–602.

Gass, S. M., and Mackey, A. J. (2017). *Stimulated recall methodology in second language research*, 2nd ed., New York, NY: Routledge.

Hayes, A. F., and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data, *Communication methods and measures*, 1(1), pp 77–89.

Iwashita, N., and Vasquez, C. (2015). An examination of discourse competence at different proficiency levels in IELTS Speaking Part 2, *IELTS Research Reports Online Series 2015/5*, https://www.ielts.org/-/media/research-reports/ielts_online_rr_2015-5.ashx British Council, Cambridge Assessment English and IDP: IELTS Australia.

Iwashita, N., May, L., and Moore, P. (2017). Features of discourse and lexical richness at different performance levels in the Aptis speaking test, *ARAGs Research Reports Online 2017/2*. British Council: London.
https://www.britishcouncil.org/sites/default/files/iwashita_et_al_layout_1_revised.pdf

Lazaraton, A. (1998). *An analysis of differences in linguistic features of candidates at different levels of the IELTS Speaking Test.* Report prepared for the EFL Division, University of Cambridge Local Examinations Syndicate. Cambridge,

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge University Press: Cambridge.

May, L. (2011). *Interaction in a paired speaking test: The rater's perspective*. Peter Lang: Frankfurt.

Nakatsuhara, F. (2012). The relationship between test-takers' listening proficiency and their performance on the IELTS Speaking Test. In L. Taylor and C. J. Weir (Eds), IELTS Collected Papers 2: Research in reading and listening assessment. *Studies in Language Testing, Vol 34,* pp 519–573. Cambridge University Press: Cambridge.

Nakatsuhara, F. (2014). *A Research Report on the Development of the Test of English for Academic Purposes (TEAP) Speaking Test for Japanese University Entrants – Study 1 & Study 2*, available online at: www.eiken.or.jp/teap/group/pdf/teap_speaking_report1.pdf Eiken Foundation of Japan and University of Bedfordshire.

Nakatsuhara, F., Inoue, C., and Taylor, L. (2017). An investigation into double-marking methods: comparing live, audio and video rating of performance on the IELTS Speaking Test, *IELTS Research Reports Online Series 2017/1.* British Council, Cambridge Assessment English and IDP: IELTS Australia.

Seedhouse, P., and Harris, A. (2011). Topic development in the IELTS Speaking Test, *IELTS Research Reports, Vol 12*, pp 69–124. IDP: IELTS Australia, Canberra and British Council, London.

Seedhouse, P., Harris, A., Naeb, R., and Üstünel, E. (2014). The relationship between speaking features and band descriptors: A mixed methods study, *IELTS Research Reports Online Series, No. 2*, pp 1–30. British Council, Cambridge Assessment English and IDP: IELTS Australia.

Seedhouse, P., and Nakatsuhara, F. (2018). *The Discourse of the IELTS Speaking Test: Interactional Design and Practice*. Cambridge University Press: Cambridge.

Tavakoli, P., Nakatsuhara, F., and Hunter, A. (2017). Scoring validity of the Aptis Speaking test: investigating fluency across tasks and levels of proficiency, *ARAGs Research Reports Online, No. 7*. British Council: London.

Taylor, L. (Ed) (2011). Examining Speaking: Research and Practice in Assessing Second Language Speaking, *Studies in Language Testing, Vol 30*. Cambridge University Press: Cambridge.

van Lier, L. (1989). Reeling, writhing, drawling, stretching and fainting in coils: oral proficiency interviews as conversations, *TESOL Quarterly 23*, pp 480–508.

Weir, C. J., Vidakovic, I., and Galaczi, E. D. (2013). Measured Constructs: A History of Cambridge English Language Examinations 1913–2012, *Studies in Language Testing Vol 37*. Cambridge University Press: Cambridge.

Appendix A: Each examiner's scoring decisions

NOTE: In order to facilitate exploration of the data by other researchers, (trainee) examiners, and examiner trainers, we produced online interactive graphs which play relevant candidate talk and display examiner notes on the scoring decision.
This online resource can be accessed on this url: https://scoring-decisions.weebly.com

Figure 18: *A's decision points for Lilly demonstrating relationship between time and score*

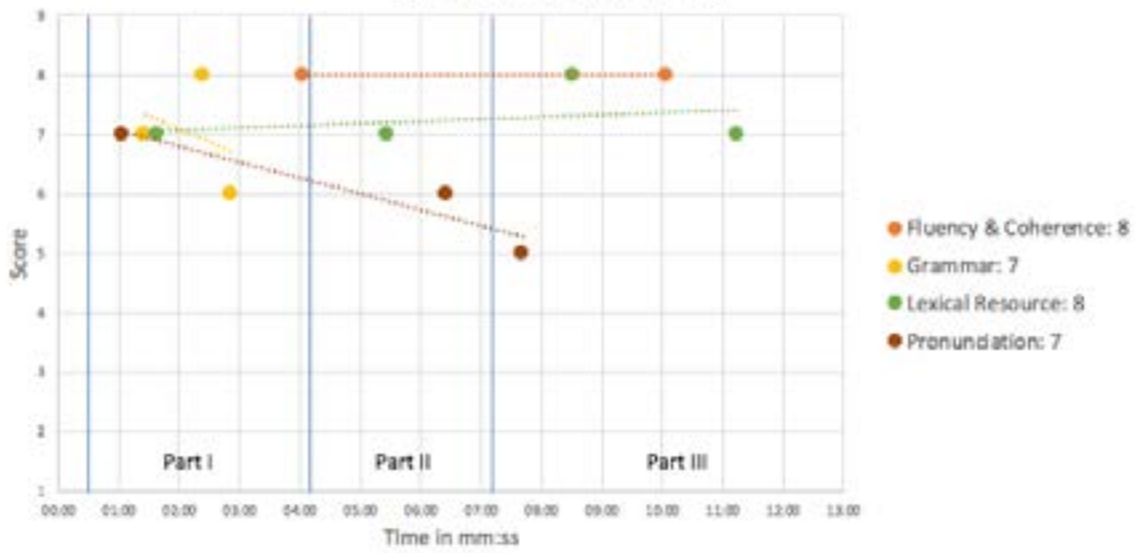


Figure 19: *B's decision points for Lilly demonstrating relationship between time and score*

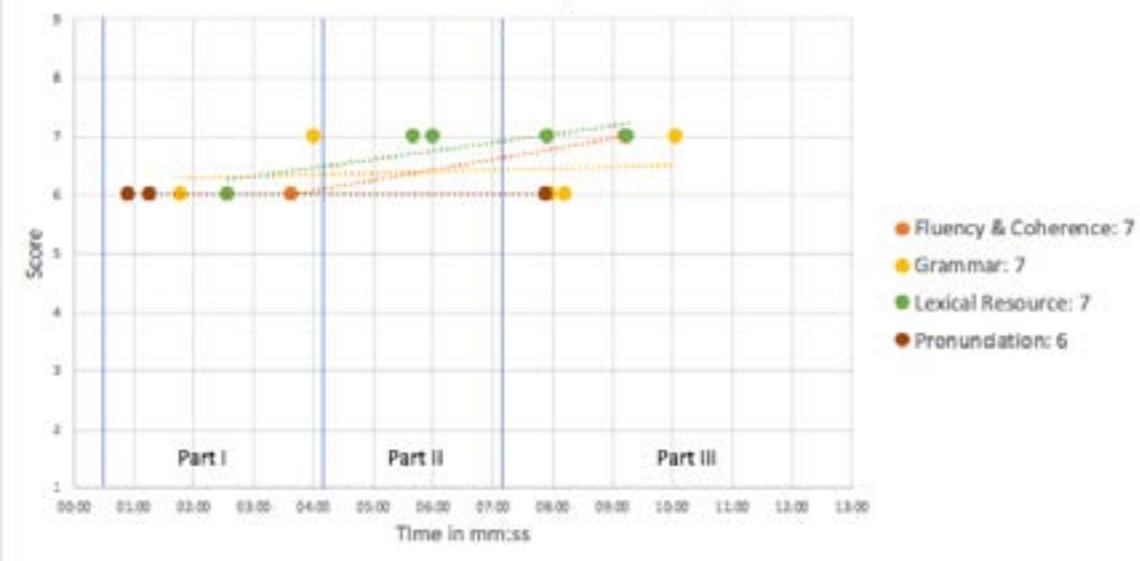


Figure 20: *C's decision points for Lilly demonstrating relationship between time and score*

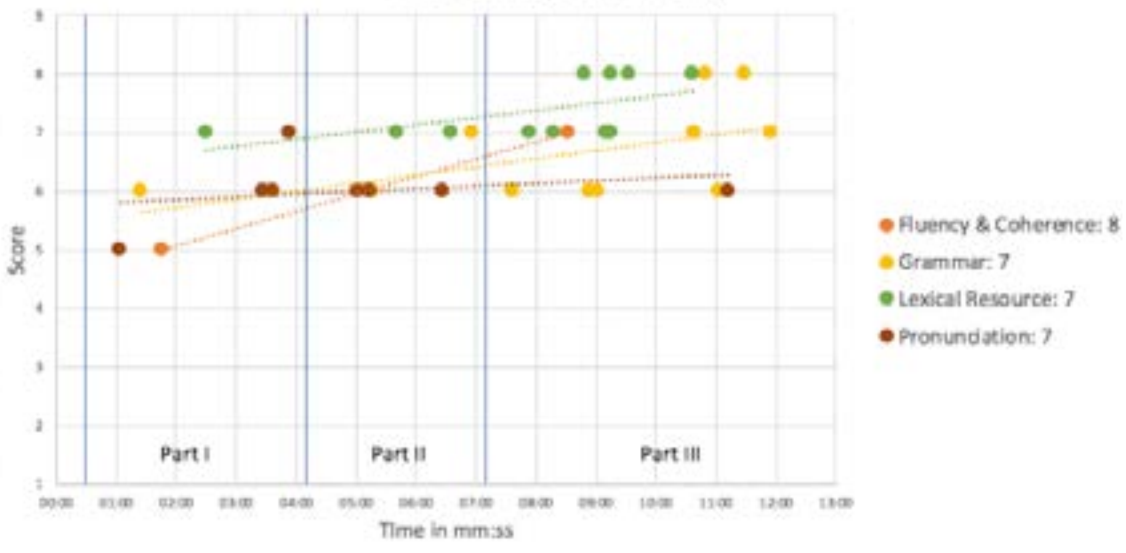


Figure 21: *D's decision points for Lilly demonstrating relationship between time and score*

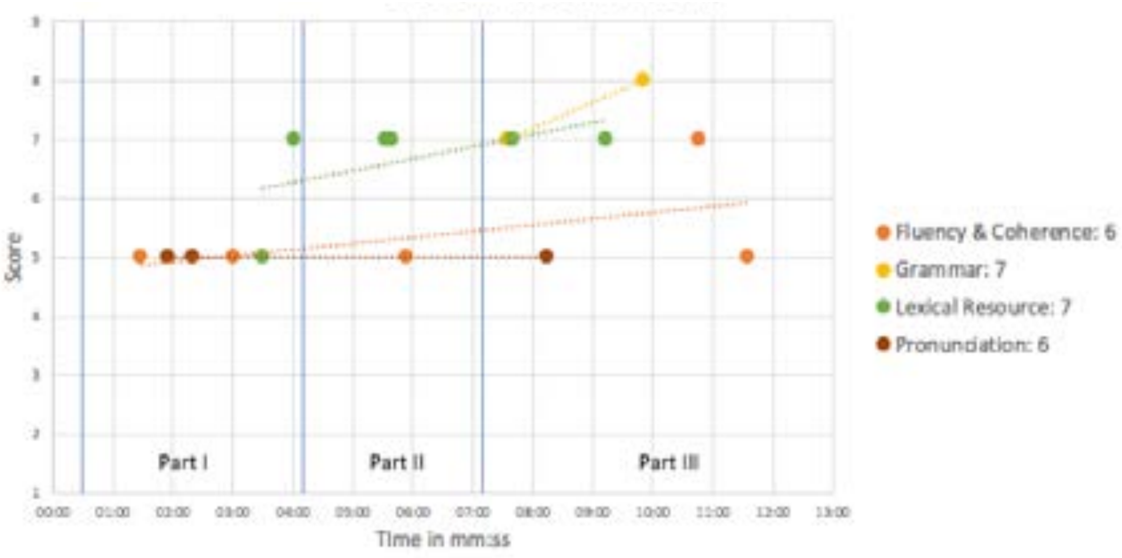


Figure 22: *A's decision points for Zoe demonstrating relationship between time and score*

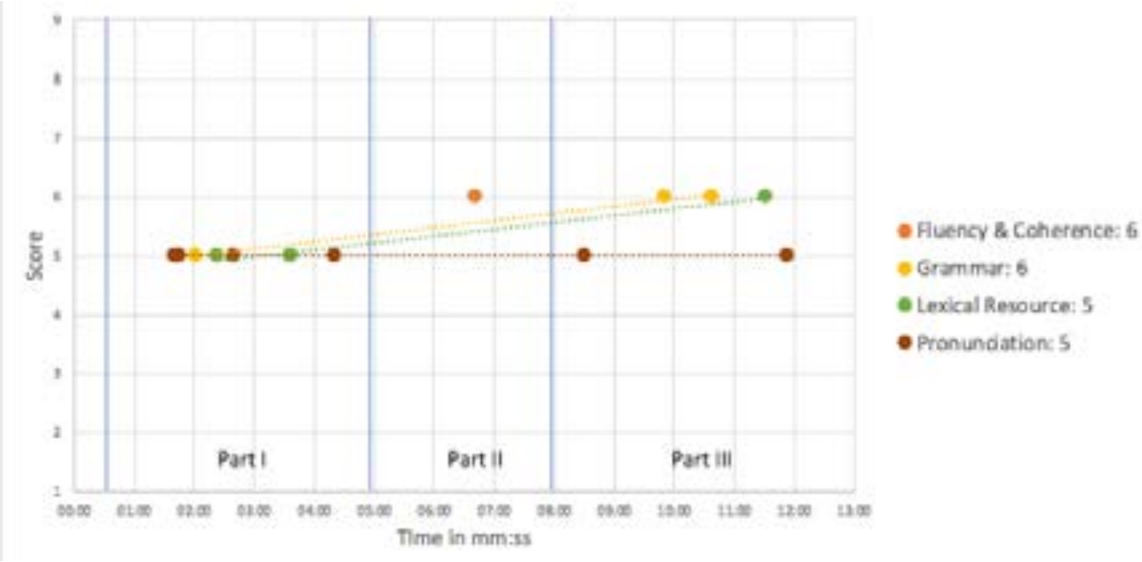


Figure 23: *B's decision points for Zoe demonstrating relationship between time and score*

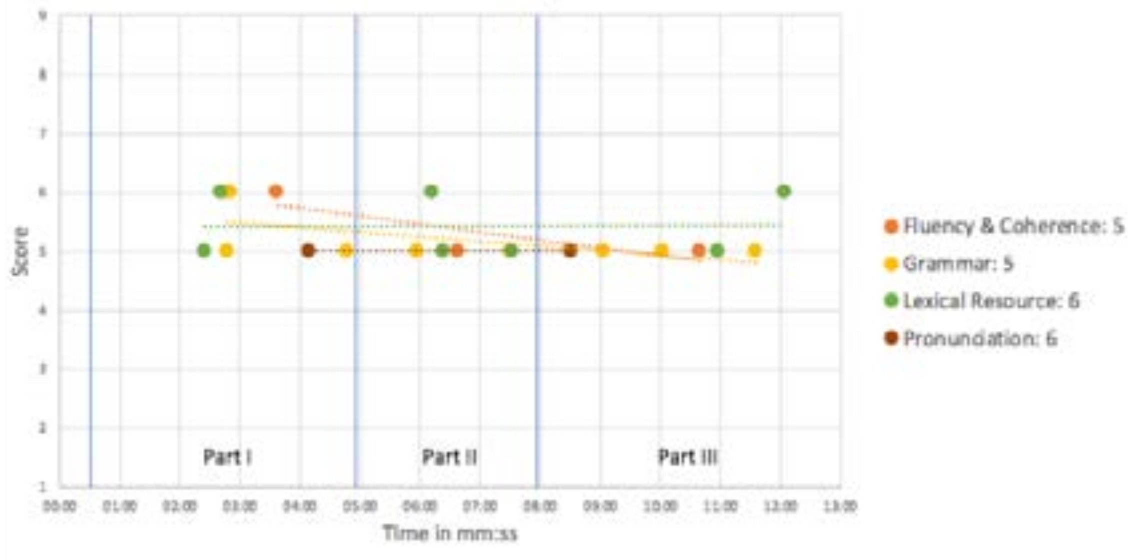


Figure 24: *C's decision points for Zoe demonstrating relationship between time and score*

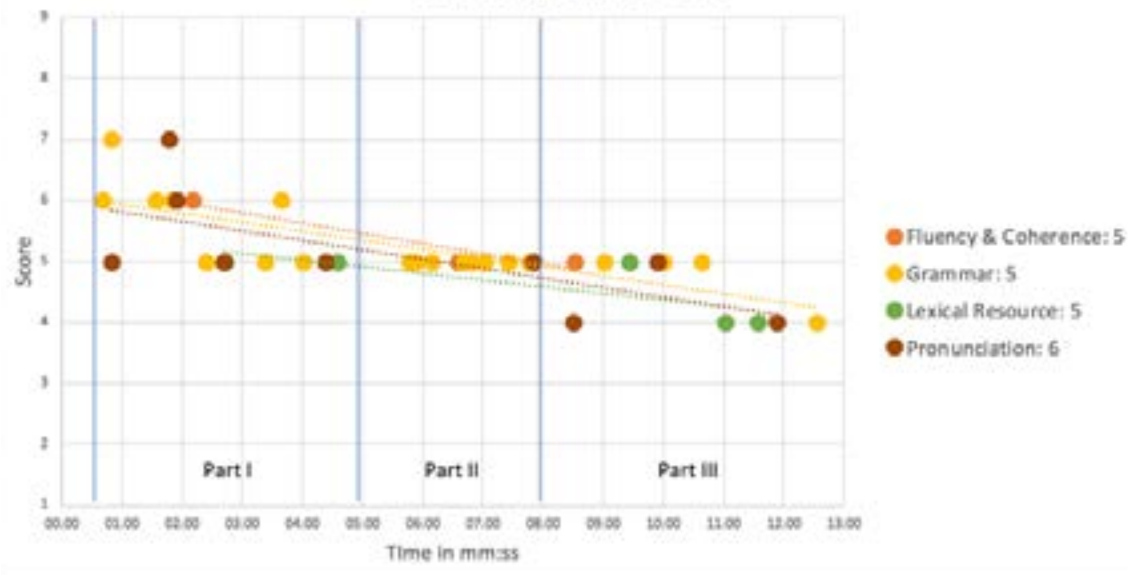
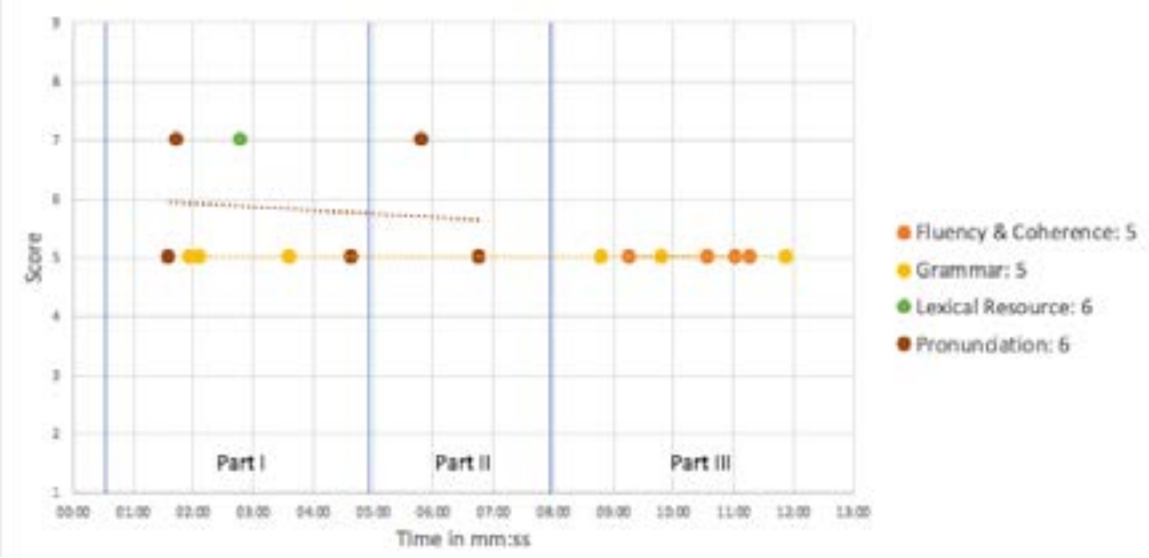


Figure 25: *D's decision points for Zoe demonstrating relationship between time and scores*


Score
Part I
Part II
Part III
00:00
01:00
02:00
03:00
04:00
05:00
06:00
07:00
08:00
09:00
10:00
11:00
12:00
13:00
Time in mm:ss
● Fluency & Coherence: 5
● Grammar: 5
● Lexical Resource: 6
● Pronunciation: 6