Raters' perceptions of topic relevance at high intermediate (CEFR B2) level in Aptis speaking performances

Judith Fairbairn

UCL, Institute of Education

Department of Culture, Communication and Media

Thesis for Doctor of Philosophy

15 September 2023

Declaration

I, Judith Fairbairn confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Thesis abstract

Deciding if a test taker's spoken response to a task is on topic is not straightforward, with likely variation in raters' perceptions (Knoch et al., 2021). There has been no dedicated focus on topic relevance in language assessment research, although this content-based criterion is often found to be a consideration. To address this gap, this study employs a mixed methods design to examine how and why raters make on-topic decisions. The research questions are twofold: to examine (1) the internal criteria that raters report using to decide if a speaking response is on- or off-topic and (2) the relationship between these internal criteria and raters' Aptis scoring and on-topic decisions with respect to rating severity, reliability, consistency and construct relevance. Thirty Aptis raters marked 45 recorded B2-level speaking responses chosen from the British Council corpus of 9000 Aptis General speaking tests. Raters provided on-topic decisions, rationales for their decisions and overall Aptis scores, and participated in semi-structured email exchanges. Many-Facet Rasch Measurement of on-topic decisions and Aptis scores investigated rater severity, leniency, fit and bias. Comparative thematic analysis of raters' written perceptions identified themes and severe, consistent and construct-relevant perceptions. Findings identified a wide spectrum of perceptions by raters of acceptable topic relevance performance and these decisions affect scoring. The findings identify areas to consider when developing topic relevance constructs and widens the socio-cognitive framework for test validation (Weir, 2005). Implications include the development of a topic relevance construct for Aptis and improving training for British Council raters. Adjustments have already been made to test specifications for item writers based on these findings.

Impact Statement

The main benefit of this research are to the British Council's language testing programme. I have worked in language testing for the British Council since 2003 in a variety of roles (research, operational delivery, test development, test production, test platform management and rating) and am now in a senior role. My expertise and ability to implement the findings from this research will have an impact on the quality of the British Council's international tests. I have already implemented findings from this research by improving task specifications and I plan to work with the rating team to implement enhanced training to the examiner pool. I will continue to mentor my colleagues and apprentices to further develop the quality of the testing team. My research and the quality of my articles will be improved because of the skills I have acquired in this PhD. I will be able to speak intelligently with academics in my line of work and be better able to influence education ministers and other government officials on country testing plans because of my expertise.

The wider testing community will benefit from this research as topic relevance scoring is an under-researched area and this project extends this knowledge. My contribution to an understanding of the construct of topic relevance can be useful for future scholarship of human rating and machine scoring. I can share all of my data with other researchers for further exploration and collaboration. I plan to write at least one journal article on my thesis and further explore the data myself.

Acknowledgements

This PhD is dedicated to Dr Jamie Dunlea. He encouraged me to do this PhD and has been a mentor and one of my great friends. I have loved talking with him about language testing. I will never forget how British Council colleagues queried how we could talk about language testing for 5-hour meetings in a Spring Gardens meeting room without a break for the toilet or food, and how colleagues regularly knocked on the meeting room doors to tell us to be quiet because we are both normally very loud and get louder when we discuss a controversial language testing issue. The man is a genius and these were some of the best times of my life.

Professor Barry O'Sullivan got me into language testing with the development of the Aptis test, which we started working on in 2008. He believed in my potential and has consistently over the years given me exceptional mentoring advice. I am very grateful for this wonderful career opportunity and his friendship.

Dr Talia Issacs was a perfect supervisor for me. I thought I could write before I met Talia, but her mentoring has moved me to another level. She always gave precise, useful and thought-provoking advice over the years. I have learned to think deeply, methodologically and critically under her supervision. She is a kind person with high expectations of quality, which I completely respect. Talia is one of the top language testers in the world and I am so incredibly grateful to have had this time with her.

Dr Ana Pellicer Sanchez was a wonderful second supervisor. She helped me craft my thesis for a non-testing expert. Her support was consistently thoughtful, extremely useful and I enjoyed our discussions.

I must thank Lucy Adam who did my second coding and one of my pilots. She is a true academic and her analysis was exceptional. I loved our discussions and cannot wait to implement the findings with her.

I also thank the raters and senior raters who participated in the study for being so professional and open with their views. They are a great team and I learned a lot from them.

A huge thank you also to Dr Johnathan Jones. We overlapped in our PhDs with Talia as supervisor. Our laps around Russell Square during Covid is something I will never forget. A PhD can be lonely and it was great to be able talk to someone who understood language testing, the Institute of Education at UCL requirements and be able to share experiences and solutions to various issues.

I thank my husband Ian Felton, who has listened to everything over dinner on many evenings. He must by now be a minor expert in speaking on-topic. I loved discussing how speaking on-topic is relevant to other fields such as politics, diplomacy and international relations.

Finally, I thank my parents, Eileen Braidwood and Bill Fairbairn. They will be astonished by this achievement. My parents taught me that the most important thing in life is to be happy and that traditional routes to success are not for everyone. I learned to be practical from my mother and tenacious from my father. I learned to be politically aware from both my parents.

It amazes me that I have become a testing expert. I would like to continue my career in ensuring fairness in global testing.

Table of Contents

RATERS' PERCEPTIONS OF TOPIC RELEVANCE AT HIGH INTERMEDIATE (CEFR B2) LEVEL IN APTIS SPEAKING PERFORMANCES 1

DECLARATION 2

THESIS ABSTRACT 3

IMPACT STATEMENT 4

ACKNOWLEDGEMENTS 5

TABLE OF CONTENTS 7

GLOSSARY FOR THE PURPOSE OF MY RESEARCH 14

1 INTRODUCTION 18

1.1 Goals of thesis 20

1.2 Research context 25

1.2.1 Socio-cognitive framework 26

1.2.2 The Aptis Speaking test 31

1.2.3 Aptis B2-level speaking task 34

1.2.4 Yes/no binary topic relevance decision in the Aptis Speaking test 38

1.3 Speaking corpus, ethics approval and data protection 41

2 LITERATURE REVIEW 44

2.1 Task topics 45

2.1.1 History of topics 45

2.1.2 Test-takers and topics 50

2.1.3 Raters and topics 57

2.1.4 Task complexity 61

2.2 Topic relevance construct and rating scales 65

2.2.1 IELTS 66

2.2.2 TOEFL iBT 69
2.2.3 Aptis 71
2.2.4 Occupational English Test 75

2.3 Rater variability when making topic relevance decisions 77
2.3.1 Rater reliability 77
2.3.2 Rater construct relevance 79
2.3.3 Writing tests 84
2.3.4 Integrated tests 92
2.3.5 Rater cognition 101

2.4 Research questions 104

3 METHODOLOGY 106

3.1 Research design 106

3.2 Participants 110

3.3 Informed consent 115

3.4 Pilot studies to inform the main study 116
3.4.1 Pilot study 1 117
3.4.1.1 Selecting responses for pilot study 1 118
3.4.1.2 On-topic rating perceptions 118
3.4.1.3 Video conference and face-to-face interview 122
3.4.1.4 Pilot study 1 findings 123
3.4.2 Pilot study 2 125
3.4.2.1 Selecting responses for pilot study 2 126
3.4.2.2 Group interview 126
3.4.2.3 Pilot study 2 findings 127
3.4.3 Pilot study 3 129
3.4.3.1 Participants 129
3.4.3.2 On-topic perceptions research instrument 130
3.4.3.3 Semi-structured email exchange 131
3.4.3.4 Pilot study 3 findings 132
3.4.4 Pilot studies summary and conclusion 135

3.5 Data collection procedures for the main study 137
3.5.1 Quantitative data collection 137
3.5.2 Qualitative data collection 139

3.6 Data analysis procedures for the main study 141
3.6.1 Quantitative analysis procedures 142
3.6.2 Qualitative analysis procedures 146
3.6.2.1 Coding rater perceptions of test-taker responses 147
3.6.2.2 Second coding 153
3.6.2.3 Thematic analysis 154
3.6.2.4 Internal criteria raters use to make on-topic decisions 155
3.6.2.5 Rater analysis 155
3.6.2.6 Qualitative analysis procedures conclusion 158
3.6.3 Mixed methods interpretation procedures 159
3.6.3.1 Raters 160
3.6.3.2 Test-taker responses 160

4 RESULTS 162

4.1 Quantitative analysis results 163
4.1.1 Aptis score analysis 166
4.1.1.1 Aptis score: raters 169
4.1.1.2 Aptis score: test-taker responses 173
4.1.2 On-topic decision analysis (separate prompts) 178
4.1.2.1 On-topic decisions: raters (separate prompts) 181
4.1.2.2 On-topic decisions: test-taker responses (separate prompts) 183
4.1.3 On-topic decisions (combined prompts) 191
4.1.3.1 On-topic decisions: raters (combined prompts) 193
4.1.3.2 On-topic decisions: test-taker responses (combined prompts) 195
4.1.4 On-topic decision analysis by prompt 201
4.1.5 On-topic decision analysis by version 207
4.1.6 Bias interaction analysis 212
4.1.7 Quantitative analysis conclusion 218

4.2 Qualitative analysis results 220
4.2.1 Coding rater perceptions of test-taker responses 220
4.2.2 Second coding 223
4.2.3 Thematic analysis 224
4.2.3.1 Generic criteria (RQ 1.1) 224
4.2.3.2 Task prompt type (RQ 1.2) 236
4.2.3.3 Task version (RQ 1.3) 242
4.2.3.4 Rating scale (RQ 1.4) 246
4.2.4 Internal criteria raters use to make on-topic decisions 250
4.2.5 Rater analysis 252

4.2.5.1 Rater severity 252
4.2.5.2 Rater consistency 254
4.2.5.3 Rater construct relevance 256

4.3 Mixed methods results 259
4.3.1 Raters 260
4.3.1.1 Lenient raters 261
4.3.1.2 Severe raters 272
4.3.1.3 Mixed profile raters 283
4.3.2 Test-taker responses 294

5 DISCUSSION 311

5.1 Rater internal criteria to decide if a speaking response is on-topic 315
5.1.1 Topic development 315
5.1.2 Question prompt content 318
5.1.2.1 Grammar expectations 319
5.1.2.2 Pictures 321
5.1.2.3 Feelings 322
5.1.3 Vocabulary 323
5.1.4 Linguistic characteristics 325
5.1.5 Approaches to collecting evidence 326

5.2 Theoretical contributions to knowledge 328
5.2.1 Aptis speaking on-topic construct 328
5.2.1.1 Aptis speaking task construct definition 329
5.2.1.2 Rater guidance and support materials 335
5.2.2 Topic relevance construct 339

5.3 Methodological contributions to knowledge 342

5.4 Limitations 343

5.5 Future research 345

5.6 Conclusion 348

REFERENCES 351

Appendix 1: Aptis General Speaking test 374

Appendix 2: Aptis Speaking B2 task versions used in this research 375

Appendix 3: Aptis Speaking Task 4 specifications 378

Appendix 4: Speaking B2 rating scale 379

Appendix 5: UCL ethics approval 380

Appendix 6: Email to raters 381

Appendix 7: Participant information sheet 382

Appendix 8: Consent form for participants 389

Appendix 9: Participant background questionnaire 393

Appendix 10: Data collection email instructions 394

Appendix 11: Sample for pilot studies 396

Appendix 12: Semi-structured email exchange – email #1 397

Appendix 13: Semi-structured email exchange – final email 399

Appendix 14: Example of email exchange questions 400

Appendix 15: Qualitative analysis codes 401

Appendix 16: Information sheet for the second coder 402

Tables

Table 1: Participant education 113

Table 2: Participant experience marking Aptis tests 114

Table 3: Template for raters to identify on-topic responses 119

Table 4: Updated template for pilot study 3 data collection 130

Table 5: Overview of pilot studies 135

Table 6: On-topic perceptions and Aptis scoring data collection instrument 137

Table 7: Example of rater analysis 156

Table 8: Qualitative data analysis procedures 158

Table 9: Overview of data analysis 162

Table 10: Observation count of topic relevance decisions for prompts 165

Table 11: Observations count of topic relevance decisions for versions 166
Table 12: Raters: Aptis score and measure table 170
Table 13: Test-taker responses: Aptis score and measure table 174
Table 14: Aptis scores: unexpected observations 178
Table 15: Raters: on-topic score and measure table (separate prompts)..... 182
Table 16: Test-taker responses: perfect on-topic agreement (separate prompts).. 184
Table 17: Test-taker responses: some on-topic disagreement (separate prompts) 185
Table 18: Test-taker responses: more on-topic disagreement (separate prompts) 186
Table 19: Test-taker responses: most on-topic disagreement (separate prompts) 187
Table 20: Unexpected on-topic observations (separate prompts) 190
Table 21: Excerpt from the on-topic prompts data file for 3-facet analysis 193
Table 22: Raters: on-topic score and measure table (combined prompts) 194
Table 23: Test-taker responses: score and measure table (combined prompts).... 196
Table 24: Prompts: on-topic score and measure table (combined prompts) 197
Table 25: Unexpected on-topic observations (combined prompts) 200
Table 26: Raters: on-topic score and measure table (prompt 1) 202
Table 27: Unexpected observations (prompt 1) 203
Table 28: Raters: on-topic score and measure table (prompt 2) 204
Table 29: Unexpected observations: prompt 2 205
Table 30: Raters: on-topic score and measure table (prompt 3) 206
Table 31: Unexpected observations (prompt 3) 206
Table 32: Raters: on-topic score and measure table (version 1) 207
Table 33: Unexpected observations (version 1) 208
Table 34: Raters: on-topic score and measure table (version 2) 209
Table 35: Unexpected observations (version 2) 210
Table 36: Raters on-topic score and measure table (version 3) 211
Table 37: Unexpected observations (version 3) 212
Table 38: Bias analysis: Raters 06-20 215
Table 39: Bias analysis: Raters 21-36 216
Table 40: Bias interaction t-statistics for Aptis score and prompts 217
Table 41: Quantitative analysis overview of raters 218
Table 42: Quantitative analysis overview of test-taker responses 219
Table 43: Exact intercoder agreement 223
Table 44: Amount of response code frequency count 225

Table 45: Clarity of response frequency count 227
Table 46: Multiple activities frequency count 236
Table 47: Rater categories 260
Table 48: Topic relevance construct and the socio-cognitive model 340

Figures

Figure 1: Aptis speaking test scoring process 40
Figure 2: Research design 109
Figure 3: Categories in second round of coding 150
Figure 4: Aptis score observation count 164
Figure 5: On-topic observation count 165
Figure 6: Aptis scoring specification file 167
Figure 7: Aptis separation and chi-squared statistics 173
Figure 8: Aptis scores Wright map 176
Figure 9: On-topic decisions specification file (separate prompts) 180
Figure 10: On-topic test-taker response labels (separate prompts) 180
Figure 11: On-topic strata and chi-squared statistics (separate prompts) 183
Figure 12: On-topic decisions Wright map (separate prompts) 189
Figure 13: On-topic decisions specification file (combined prompts) 192
Figure 14: On-topic strata and chi-square (combined prompts) 195
Figure 15: On-topic decisions Wright map (combined prompts) 198
Figure 16: Bias interaction specification file 214
Figure 17: Coding rater perceptions of test-taker responses (round 2) 221
Figure 18: Memos while coding 221
Figure 19: Spectrum of on-topic perceptions 329

Glossary for the purpose of my research

Affective schemata: Positive or negative emotional feelings or reactions test-takers have about a topic.

Construct: An ability or mental trait which cannot necessarily be directly observed or measured.

Construct irrelevance: Occurs when raters systematically use other criteria not in the rating scale to make decisions and allocate scores.

Construct relevance: Occurs when raters systematically make decisions and allocate scores as expected and intended by the test developers.

Criterion rating scales: A scale which measures test-takers' performance using a level of achievement or objectives.

Dichotomous scoring: Scoring based on two categories (e.g., yes/no).

Examiner: Conducts a test as an interlocutor by reading out the question prompts and also scores the performance. Also called a rater or marker.

FACETS: A computer programme which facilitates Many-Facet Rasch Measurement to determine if raters are consistently making the same types of scoring decisions and to identify problematic tasks or criteria in rating scales.

Fit: The extent of agreement in the scoring between individual raters and the sample of raters. Used to look for inconsistent or overly consistent scoring.

Holistic rating: Judging overall spoken performance to produce one score.

Independent speaking task: The question prompt is short and test-takers need to use their own language to respond.

Interlocutor: A person who asks questions and interacts with the test-taker to elicit spoken language.

Inter-rater reliability: The degree of scoring agreement between raters on the same sample of spoken performance.

Intra-rater reliability: The degree to which the same rater scores the same spoken performance similarly on different occasions.

L1: A test-taker's first or main language.

Lexis: A term used to refer to vocabulary.

Many-Facet Rasch Measurement (MFRM): Uses response probabilities to determine if raters are consistently making the same types of scoring decisions and to identify problematic tasks or criteria in rating scales

Observed score: The score obtained by a test-taker.

Polytomous scoring: Scoring based on at least three categories (e.g., 'on topic', 'borderline on topic' and 'off topic'.

Proficiency language test: A test which measures general language ability or skill over a wide range of functions and topics at any given level and is not based on a specific curriculum.

Question prompt: Tasks designed to elicit test-taker spoken performance.

Rater: Someone who assigns a score to a test-taker's spoken response. Also called a marker or examiner.

Rating: Assigning a score to a test-taker's performance using judgement.

Rating scale: Categories used for making judgements of spoken performance.

Reliability: The rating consistency or stability.

Response: The spoken performance by a test taker.

Semi-direct test: Test-takers record responses on a digital device (e.g., laptop or tablet) to pre-recorded prompts. There is no human interviewer.

Socio-cognitive framework: Collects test validation evidence around the test-taker, the test system, and the scoring system to show that the quality of scores provides adequate, fair and appropriate inferences about test-takers' abilities for a given context and consequence of test use.

Specifications (test or task): Documentation which gives details of the characteristics of the input and expected test-taker responses, and all other information required for item writers to produce new versions (forms) of the task question prompts.

Task: A question prompt (or prompts) that elicits a test-taker's spoken response.

Task type: A type of question prompt that elicits a particular type of response. For example, a speaking task type could be a question prompt that elicits a spoken response of 30 seconds in the past tense about oneself.

Validation/validity: The extent to which scoring decisions allow inferences to be made about test-taker spoken proficiency which are appropriate, meaningful and useful for a specific test purpose. An ongoing process of justifying particular contexts of test use.

Version: Alternate or parallel forms of the same test or task, which have the same average mean difficulty.

1 Introduction

Deciding if a test-taker's spoken response is topic relevant, on the surface, may appear simple. In practice it is not so straightforward and there is likely to be a variation in how raters perceive if a response is on-topic. It has been my impression that raters find it difficult to decide if a response is on-topic from my experience of managing raters for two international tests. Raters seem to struggle with knowing the threshold for when a test-taker response has topic relevance.

My practical expertise from 20 years in International English Language Testing System (IELTS) and Aptis rater management and research roles through my employment with the British Council provided personal familiarity with the context and access to a pool of raters to conduct the current study. My roles have included managing and training raters; managing test development, test delivery and operations; managing project teams; and conducting test validation work for the speaking and writing skills. I was also part of the team that developed the test used in this research and have had a personal mission to enhance the quality of this test.

After doing this research I am even more convinced that colleagues need support in this area. I have been responsible for the teams that write and mark the Aptis tests and have been able to implement some of these findings into the areas of item writing and rater training, as well as create a research agenda for the future.

There are real life implications if raters do not understand how to rate topic relevance. Test-takers might attempt to shoehorn a memorised response into a speaking test, or they may use strategies that continuously redirect their performance towards something rehearsed (Fernandez, 2018). Research shows that if raters perceive that a spoken response is memorised and off-topic, they will score

the test-taker at a lower level (Burton, 2018; Cox, 2017; Gates, 2018). This research might help to extend these findings.

There are a variety of reasons why raters may struggle to mark topic relevance. It is well-known in the literature that when rating overall spoken language, raters vary in their perceptions of successful spoken performance. Examiners listening to test-takers' spoken performance do not hear the same message (Bygate, 1987; Douglas, 1994). Raters can find it difficult to remember the salient features of a test-taker's response or have bias towards certain rating criteria. Raters may also vary in their tolerance for perceived spoken errors (Canale & Swain, 1980). Various rater characteristics, such as background and experience, can also lead to rating differences (Douglas, 1994; Knoch et al., 2021). These rating differences for overall spoken production are also likely to be present when marking topic relevance criteria, although, to my knowledge, there is not much research in this area.

In addition to individual and systematic rating differences, there is complexity in the interaction between the test takers, raters, speaking tasks and rating scales. The type of task and rating scale, as well as test-taker characteristics and background knowledge, can interact and affect language production and scoring (Bachman & Palmer, 1996; Khabbazbashi, 2021). This complexity in testing the speaking skill has been thoroughly studied (Davis, 2016; Isaacs, 2016; Isaacs & Thomson, 2020; Taylor & Galaczi, 2011) and is further discussed in the literature review with respect to topic relevance.

The research is extensive on scoring variability for the speaking skill but has not fully extended to the conception of topic relevance, as far as I am aware. There is a limited understanding of how and why raters decide if a spoken response is on-topic.

It is mostly unknown how the task prompts, test-taker responses and rating scales affect topic relevance decisions made by raters. The research is more thorough for the writing skill, which has some useful learning for the speaking skill, and is discussed in the literature review.

This thesis aims to contribute to knowledge with a fuller understanding of the topic relevance construct in speaking tests. A construct is a test-taker competence which cannot necessarily be directly measured but is a meaningful interpretation of observed behaviour (Chapelle, 1999). A rater's topic relevance decision is the observed behaviour. Fulcher and Davidson (2007) describe a construct as a description of the components in a model of communication. Similarly, for Bachman and Palmer (1981), a construct is a theory of language traits. Weir et al. (2013) go further and include the interactions between the cognitive abilities, context, scoring criteria and interpretation of the results in a construct definition. This thesis will examine one model which measures topic relevance and investigate the observed rater behaviour to determine if this approach to scoring topic relevance leads to meaningful interpretations of what it means to speak on-topic.

1.1 Goals of thesis

The first goal of this thesis is to identify features in the speaking tasks used in the current study which may affect topic relevance rating. There are three example tasks and each task has three question prompts. There is scope to identify elements in the question prompts which cause raters to mark topic relevance more severely, inconsistently or with bias. This exploration has implications for task type design and item writing.

The second goal is to start to develop a construct definition of topic relevance for the Aptis task used in this research (see section 1.2 for a description of the Aptis test, the task used in this research and the rating scale). Aptis currently does not have a clear construct definition of what it means to speak on-topic for the task used in this research. Examiners decide if the spoken response for each question prompt is on-topic or off-topic. It is a yes or no decision without any criteria for making this decision. There is no information on what it means to speak on-topic. The current study will investigate how and why raters make topic relevant decisions when marking this task by identifying internal criteria raters use to make decisions about topic relevance.

The third goal is to contribute to the theoretical understanding of the topic relevance construct in speaking tests more generally. The focus is on the socio-cognitive framework (Weir, 2005), which guides test developers on how to establish scoring validity evidence with respect to topic relevance but does not include raters' perceptions of test-takers' spoken responses. The framework discusses rater reliability in general but does not specifically discuss how raters deal with topic relevance decisions made by test-takers. The focus also is mostly on spoken interaction with an interlocutor. There is scope from my research to widen the socio-cognitive framework to include rater perceptions of topic relevant decisions made by test-takers in a semi-direct speaking test.

The final aim is to develop better rater guidance and support materials that explain the speaking topic relevance construct to raters. Aptis raters regularly ask senior raters and managers for advice on how to make the yes/no binary topic relevance decision so there is a lack of clarity in marking this part of the test. Researchers in

the testing field agree that raters must understand the principles behind the rating scales and consistently interpret the rating scale descriptors. This is done through training and standardisation (Alderson et al., 1995; Davis, 2016; Elder et al., 2007; Hamp-Lyons, 1995; Knoch et al., 2007; Lim, 2011; Shaw, 2002; Weigle, 1994, 1998, 2002; Wigglesworth, 1993; Xi & Mollaun, 2011). Davis (2016), for example, found a clear correlation in the use of example test-taker responses and rating accuracy. The British Council could develop guidelines to help raters identify if a response is on-topic, with a selection of clear-cut benchmark speaking performances from this research thus improving the operationalisation of the yes/no binary topic relevance decision in the Aptis rating scale.

This mixed methods study aims to contribute to the topic relevance decision-making discussion by analysing rater perceptions, based on their accounts, on whether spoken responses are on-topic or off-topic. This project uses test-taker responses and focuses on rater decision-making processes and strategies. This type of rater perception research has been conducted for the written productive skill (e.g., Baker, 2012; Kuiken & Vedder, 2014; Lumley, 2005), but there is little research on the speaking skill and none specifically, as far as I am aware, on how raters decide how and why a test-taker's response is on-topic.

This thesis consists of five chapters. The rest of this introductory chapter 1 explains the research context and describes the Aptis test, the test task used in this research and the rating scale. It includes a discussion on the socio-cognitive model used to develop the Aptis test and how the rating scale incorporates a yes/no binary topic relevance decision. It includes information about the British Council speaking corpus used in this research. I also document the steps taken to obtain data protection

agreement to use the corpus of spoken responses as well as ethical matters raised during ethics approval.

Chapter 2 is the literature review, organised into four sections. The first section includes the literature about task topics. I discuss the history of topics and examine how topic knowledge and task complexity can affect test-taker performance. The gap in understanding is if and how this topic knowledge and task complexity also affects rater topic relevance decisions. The second section investigates the topic relevance construct and how speaking on-topic in language tests has been conceptualised, defined, measured and interpreted in the literature and by testing organisations in their rating scales. The third section reviews the literature on rater variability as quantitative and qualitative concepts. I discuss rater reliability and measurement error, rater construct relevance and key concepts from language testing judgement, cognition and decision-making research that are relevant to this study. This section also includes research from writing tests and integrated testing that have a larger body of literature on rating topic relevance. This section concludes with the research questions.

Chapter 3 outlines the mixed-methods research design and methodology. This chapter includes information about the participants, informed consent, data collection instruments and data analysis procedures used in this thesis. This chapter also includes the three pilot studies, which informed the methodology for the main study. I include reasons for using Many-Facet Rasch Measurement for the quantitative analysis and comparative thematic analysis for the qualitative analysis and why I chose email exchanges to collect additional follow-up qualitative data.

Chapter 4 includes the quantitative and qualitative findings and the mixed methods integration, interpretation and triangulation of the entire data set. This chapter is in three parts. The first two parts are the quantitative and qualitative analyses, which I conducted separately and in parallel. The third part is the mixed methods interpretation which integrates the quantitative and qualitative findings into a deeper triangulation and interpretation.

Building on insights from the previous chapter, Chapter 5 is a discussion of the main findings from the research with a link to the literature, the aims of the research and the research questions. It includes a discussion on the internal criteria raters use to decide if a speaking response is on-topic and features in the speaking tasks used in the current study which may affect topic relevance rating. There is a discussion on the theoretical and methodological contributions to knowledge arising from this research including what elements could be included in a topic relevance construct for this Aptis task and speaking on-topic more generally, suggestions for enhancements to the socio-cognitive framework and using write-aloud protocols and email exchanges as research methods. The chapter includes suggestions for better rater guidance and support materials that explain the speaking topic relevance construct to raters. The chapter concludes with limitations of the study and suggestions for future research.

This chapter so far has introduced the focus of the study: the topic relevance decisions raters make when marking speaking tests. There has been an overview of rating issues and the main reasons why there can be rating variability. I also described the goals of this thesis and how the thesis will develop. The key point to note in this introduction is that there is a gap in understanding in how and why raters

make topic relevance decisions. The next section outlines the background and context of this present study.

1.2 Research context

To address the central concern of this thesis of how and why raters decide if a test-taker response is on-topic, I used one test task from the British Council's Aptis General speaking test (O'Sullivan et al., 2020). The task I used includes on-topic marking requirement in the rating scale, which has provided the opportunity to investigate this construct.

The British Council developed the Aptis General test in 2012. Institutions and businesses use the test to assess adults for a variety of English general proficiency purposes such as placement into classes and employment recruitment decisions. Increasingly, Aptis is being used for high stakes decisions such as English proficiency tests for university graduation in Mexico (British Council, 2023a) and as a university exit test in Spain recognised by the Office of Qualifications and Examinations Regulation in England (Ofqual) (British Council, 2023c).

The test includes listening, reading, writing and speaking tests plus a core grammar and vocabulary test. Test-takers can take one component or all components. Numerical scores for all components are reported on a scale of 0–50. The numerical scores for the listening, reading, writing and speaking tests are converted to levels on the Common European Framework of Reference (CEFR) (Council of Europe, 2001, 2018) using the core grammar and vocabulary results to fine-tune decisions. The British Council awards a final Aptis score out of 200 and an overall CEFR level if all components are taken. See Dunn (2024) for a detailed overview of Aptis scoring.

This thesis focuses on one task in the speaking test. The following sections discuss the framework used to develop and validate the Aptis test, the Aptis speaking test and the task used in this research, as well as the topic relevance rating requirement for the task.

1.2.1 Socio-cognitive framework

The British Council designed the Aptis test using an adaptation of the socio-cognitive theoretical framework of test development and validation originally developed by Weir (2005) and further expanded by O'Sullivan & Weir (2011) and others. The framework focusses on testing cognitive processing abilities and highlights that language is not purely a linguistic activity but also a social phenomenon (Weir et al., 2013). The framework describes concrete features of a test and test-taker performance and expects researchers to collect supporting evidence and reflect on the construct conceptualisation. The evidence needs to show that the quality of scores provides adequate, fair and appropriate inferences about test-takers' abilities for a given purpose or test use (Canale, 1987; Chalhoub-Deville & O'Sullivan, 2020; Knoch & Chapelle, 2018; O'Sullivan & Chalhoub-Deville, 2021).

Weir (2005) developed the socio-cognitive theoretical model to include six areas that language testers should query. The first area includes an understanding of the physical, physiological, psychological and experiential characteristics of test-takers and ensuring the test is appropriate for the test-takers. Test-taker characteristics can include age, gender, background knowledge and experience.

The second area is theory-based validity, also called cognitive validity, and involves an analysis of the meta-cognitive or mental language processing required for language use and determines if a test is targeting these executive resources

appropriately. Testers must have a theory of spoken performance and collect evidence that the test is measuring the intended construct. For example, Bachman's (1990) communicative language ability model includes definitions of language competence, strategic competence and psycho-physiological mechanisms. Levelt's (1990) model of the executive process is another approach to define the theory of spoken performance. The cognitive processes needed to perform the test should be the same as those used in real life, as defined by the tester's theory of spoken performance (Field, 2011; O'Sullivan & Weir, 2011).

Context validity is the third area and investigates if the linguistic demands of the task are appropriate and generalisable. Research includes a review of test content, including test-taker background and subject matter knowledge. Do the test-takers have the content-related executive resources needed to accomplish the task? This area also reviews the fairness in test administration including interlocutor performance by speaking examiners. There is a close relationship between context and theory validity. The theory of the speaking model must have a clear understanding of the context of the test-takers and the test delivery model.

Scoring validity is the fourth area and includes the extent to which test results are stable, dependable, meaningful and consistent, and free from bias and measurement errors. Areas to investigate in speaking tests include parallel forms reliability and rater and rating scale reliability. There is a balance to strike between enhancing scoring validity without compromising the other validities. For example, scoring spoken responses needs to be easily understood by examiners, which requires a simple rating scale. With a simple rating scale, scoring validity could be high but the theoretical validity could be low because we are not testing the intended construct

fully enough. The context of the test also affects scoring validity, for example interlocutor performance by speaking examiners. Context, theory and scoring validity are separate parts of the model but the symbiotic interaction between these elements is the essence of the unified approach to construct validity (O'Sullivan & Weir, 2011).

The test also needs external validation, or the successful realisation of construct validity (O'Sullivan & Weir, 2011). Consequential validity is the fifth area and researches the social consequences of test interpretation: appropriateness, meaningfulness and usefulness (Messick, 1989). This type of investigation collects evidence from stakeholders on test use including washback in classrooms and other consequences of the test on the test-takers. Criterion-related validity is the final sixth area and investigates how well test scores correlate with performances on other speaking tests or other suitable external criteria of spoken performance.

The socio-cognitive framework outlines the evidence testers should collect for their validity argument and highlights the interaction between different validity evidence. The evidence is collected in two parts. Before a test is given to test-takers, the attributes of the intended test-taker, the context of the testing situation and the expected theoretical or cognitive demands of the test should be documented. Activities after the test include statistical analysis of scoring test data and consequential and criterion-related studies, although Knoch et al. (2021) argue that test developers should consider the post-test activities at different stages of the test development and validation cycle. Taylor (2011) is one researcher who used the socio-cognitive framework to investigate Cambridge speaking tests. The researcher shows the benefits of the framework and how evidence can be collected

systematically and comprehensively to validate the construct definition and operationalisation of a test.

The Aptis test used in this thesis was designed using an adaptation of the socio-cognitive framework (O'Sullivan, 2015). The adaptation focuses on collecting test validation evidence around three elements: the test-taker, the test system, and the scoring system at each step in the test development and test use stages. Test-taker characteristics and cognitive validity in the Weir (2005) model are consolidated into the test-taker category. The test system is similar to context validity. Scoring validity includes external validity, or the value of decisions.

This thesis provides a theoretical contribution to scoring validation and is part of a broader research agenda on the Aptis speaking test and scoring validation. For example, Iwashita et al. (2017) investigated features of discourse competence and vocabulary use in the Aptis speaking test. This study researched the expectations of vocabulary, coherence and cohesion at different CEFR levels. There is little reference to topic relevance and only at lower levels of performance. Kang and Johnson (2021) investigated which linguistic features of spoken performance could be extracted and used in automated assessment. The researchers identified 36 lexico-grammatical measures. Again, the focus was on the difference in linguistic performance between high- and low-level test-takers. None of the studies on the Aptis speaking test, as far as I am aware, specifically investigate topic relevance.

The focus of this thesis is to investigate one aspect of Aptis scoring validity, specifically collecting evidence around the on-topic decisions made by raters, with the aim of contributing to a definition of the Aptis speaking test construct more clearly and improving rating consistency. The approach I take is similar to Weir (2005) and

Taylor (2011), who separate scoring and external validity; external validity is out-of-scope for this study.

Weir (2005) explains the complexity of how to establish scoring validity evidence for speaking tests. The socio-cognitive validity framework includes checklists of information that should be included in speaking test specifications (e.g., which functions are being tested), parallel form investigations, rating scale development and rating reliability. The framework also includes areas that test developers need to review, which includes providing clear and explicit instructions and training to raters, agreeing acceptable responses and appropriate scores at the outset of scoring, specifying performance criteria and making the rating scale as easy to interpret as possible to reduce raters' subjective judgement.

Weir (2005) discusses content knowledge required by test-takers to complete a task as part of context validity (p. 75) but, while there is a detailed discussion on how test developers should confirm rater reliability in the scoring chapter, there is no specific discussion on how raters deal with a test-taker's approach to addressing the content in a question prompt. The framework also discusses rater reliability in general but not specifically how raters deal with topic relevance decisions made by test-takers. The focus also is mostly on spoken interaction with an interlocutor. This research study aims to widen the socio-cognitive framework to include rater perceptions of topic relevant decisions made by test-takers.

In the next research context sections are descriptions of the Aptis speaking test, the task used in this thesis and the yes/no binary topic relevance decision used in the rating scale.

1.2.2 The Aptis Speaking test

This thesis focuses on the Aptis General Speaking test (Appendix 1). The speaking test uses tasks to elicit language through the use of question prompts (Norris, 2016; Skehan, 1998). The speaking test has four tasks targeted at different levels on the Common European Framework of Reference (CEFR) (Council of Europe, 2001, 2018). The CEFR measures language proficiency at six levels: A1, A2, B1, B2, C1 and C2 with 'can-do' statements, or descriptors, for each level of proficiency. The Aptis speaking test has four tasks at A1, A2, B1 and B2 level. Each task has three question prompts and test-takers respond to all four tasks (12 questions). The scores are combined into a final score out of 50 and given a level on the CEFR. The A1 level task elicits three short responses (30 seconds each) to questions on familiar and concrete topics. The A2 task uses one picture to elicit description and opinions about the photo. The B1 task has two photos and elicits comparison and preference. The B2 task, the focus of this thesis, is a long turn (2 minutes) with 1 minute preparation and is further described in the next section. The multi-question approach with 4 speaking tasks reduces the likelihood of all topics being unfamiliar to test-takers. Also, raters mark only one task for each test-taker, which reduces rater bias.

The online Aptis speaking test is semi-direct in that test-takers record responses to pre-recorded prompts. Semi-direct tests mean that there are no humans to deliver the tests. Instead, the computer reads out the question prompts and records the test-takers' spoken performance. Raters mark the recordings after the test-taker has completed the test. These types of tests are not new (Stansfield & Kenyon, 1992a); however, they are becoming more prevalent and therefore important to study. They have the benefit of standardising test delivery, which can increase scoring reliability and fairness (Brown, 2000, 2003; de Jong et al., 2012; McNamara et al., 2002;

Nakatsuhara, 2008) and allow testing organisations to administer tests globally in remote locations. However, without a human interlocutor, there is no one to guide the test-taker if they begin to speak off-topic.

Globally dispersed raters mark the spoken performances. Raters are trained to use the rating scales using an online rating system (Knoch et al., 2015, 2016) and undergo regular standardisation training to ensure the consistent application of the rating criteria to task performances. Raters log onto the system at a time of their choosing to mark tests from any country. The British Council continuously delivers tests (24 hours a day, 365 days a year) and raters mark the tests within 48 hours. Examiner managers, senior examiners and quality assurance examiners provide technical and academic support to the examiners.

Aptis has a ‘control item’ (CI) system that has two functions: it provides operational estimates of rater reliability and identifies raters who may not be marking to standard. Integrated into the live rating system are pre-agreed scored benchmark performances described by Shaw and Weir (2007) as “gold standard seeding” (p. 307). The marking platform automatically presents raters with CIs while they mark live tests. If a rater scores a test-taker response outside of the tolerance threshold for the pre-agreed score for the CI, the system automatically suspends the rater from marking that task type. The tolerance threshold allows raters to mark differently to the benchmark scores but within an expected level of tolerance. The data also provide a means of estimating the degree of reliability between raters and the intended benchmark scores for CIs (Fairbairn, 2015b; O'Sullivan et al., 2020).

The socio-cognitive framework is operationalised in the Aptis test by using task specifications. Specifications include characteristics of the input (question prompt)

and expected test-taker responses, and all other information required for item writers to produce new versions (forms) of the task question prompts (Bachman & Palmer, 2010; Fulcher & Davidson, 2007). For test security reasons test developers produce parallel forms of tests or versions (Khalifa & Salamoura, 2011) and it is important for test fairness that versions elicit comparable speech. Specifications help ensure that test versions have very similar mean difficulties when administered to the same person. The specifications also include carefully selected task context parameters to influence the cognitive processing invoked during the test (Taylor & Galaczi, 2011). The British Council puts the new tasks through a rigorous quality assurance process including checklists, editing and proofreading before going into live tests. The specifications are publicly available for researchers and Aptis test stakeholders (O'Sullivan et al., 2020).

Shin et al. (2022) provide an overview of the Aptis test (speaking, writing, reading, listening and grammar & vocabulary). The review explains how the test can be localised (adapted) for clients including changes to the pronunciation, accent and word choice for the input (e.g., Aptis American) and that Aptis also can be delivered remotely with video monitoring or using pen and paper. The validity evidence focuses on test content, response process, internal structure, relations to other variables and consequences. Concerns about the speaking test included limited validity evidence to support the identification of C1 test takers, concern that the tests are delivered by computers and without a dialogue with an examiner (research findings were mixed), limited scoring validation evidence and a gap in direct evidence of the consequences of test score interpretations by stakeholders. Suggestions for further research for the speaking test included collecting evidence from test-takers on their response processes to check for construct-irrelevance and

other issues, identifying how the test performs in specific contexts including local validation, the impact of the computer-delivered mode and comparing how stakeholders use the total scores and the overall CEFR levels. The research discusses the rater reliability research I conducted (Fairbairn, 2015b; Fairbairn & Dunlea, 2017) and recommends additional rater reliability studies with the new rating scales. The paper does not detail any other concerns about the Aptis speaking rating scales or about topic relevance measurement.

1.2.3 Aptis B2-level speaking task

This thesis focuses only on the B2-level speaking task in the Aptis Speaking test and uses test-taker responses rated at B2-level. I chose the B2 task because it has specific on-topic rating requirements. Raters decide if test-takers answer the three questions in the task on-topic or off-topic as three binary yes/no decisions. I considered a B1 level task, which has the same on-topic rating requirement, but the prompts are too simple so there would be less divergence in on-topic decisions by B2-level test-takers. I chose B2-level test-takers because I wanted to focus on the on-topic construct and not be concerned about language level interference (e.g., grammar, pronunciation, fluency errors) from low-level test-takers struggling to respond to a B2-level task. B2-level test-takers should be able to adequately engage with the B2 task and answer the questions.

The three B2 speaking tasks used in the current study are in Appendix 2. These are tasks from Aptis operational test delivery which are no longer in use. Test-takers and teachers use these versions as practice materials and academics use them for research. I chose these tasks because they are no longer confidential and the British Council has a corpus of test-taker spoken responses to these tasks. There is a

process to request access to use this data set as well as annual funding grants to do research on the Aptis test (British Council, 2023b).

The specifications for the task used in this thesis are in Appendix 3. Test-takers are expected to integrate their responses into a long turn, with one minute to prepare a response. The photo is only for additional contextualisation of the topic and the questions do not refer directly to the photo. The specifications carefully control the vocabulary level in the question prompts using word frequency lists (Kremmel & Schmitt, 2016). The specifications also control the grammar level and topics using the British Council (2015) Core Inventory for General English.

The three question prompt types have different complexities and elicit different test-taker spoken production on three dimensions: personal vs general; present vs past tense; and concrete vs abstract topic. Personal, present and concrete target lower-level language abilities whereas general, past and abstract target higher-level language abilities (Council of Europe, 2001, 2018). The structure of this task follows an established testing approach that question prompt complexity effectively discriminates test-taker language performance output (Council of Europe, 2001).

Question prompt 1 (Tell me about an event or activity in the past) elicits description of a personal, past and concrete experience. The personal and concrete nature of these prompts elicits a B1 level response with the past tense requirement moving the prompt to a higher B1 level. Prompt 1 is meant to be the easiest prompt because it elicits concrete personal description.

Question prompt 2 (How did you feel about…) is abstract and asks test-takers to discuss their feelings about the personal experience in prompt 1. Prompt 2 elicits a personal, past and abstract response. Test-takers are still talking about themselves

and in the past tense. This prompt targets B2-level ability because test-takers are now speaking about an abstract topic (feelings).

Question prompt 3 is abstract and requires a discourse on society (e.g., Is it important for children to do outdoor activities?). Prompt 3 elicits a general and present tense response. The third prompt is meant to be the most challenging and requires the test-taker to talk about an abstract topic and not about themselves. The impersonal and abstract nature of this prompt moves this prompt into the higher B2 range.

The task has features to ensure that the question prompts are globally accessible. The question prompts appear on the screen and the computer also reads aloud the question prompts to the test-taker to ensure that the questions are widely accessible. There are B2-level speaking tasks for a range of accessibility needs (e.g., Braille, large print, lip reading, picture-free) as well as a full approach to equality, diversity and inclusion in testing for all test-takers (Fairbairn & Spiby, 2019). However, I have not included tests with accommodations in the current study. The photo and text remain on the screen as the test-taker responds to the questions. The required content knowledge to answer the questions is expected to be neutral and globally accessible.

The length and procedures for this task are standardised. Test-takers have one minute to prepare a two-minute spoken response. They can make notes with a pen and paper. Long turns with planning time tend to be more prepared (Bygate, 1987) and the planning time can affect performance (Wigglesworth, 1997). This thesis, however, is not aimed at understanding how test-takers perform under different preparation conditions but rather on how raters perceive spoken responses. It is not

clear if raters consider preparation conditions when making topic relevance decisions, and it is out of scope in the current study to specifically investigate.

Raters use a rating scale to mark this B2-level task, which has a 7-point scale (0-6) (Appendix 4). At the B2-level on the scale, the focus of the current study, the criteria for grammar, vocabulary, pronunciation, fluency and cohesion are identical for a B2.2 (Aptis score 4) and B2.1 (Aptis score 3). The only difference is the requirement for the test-taker to speak on-topic. For a B2.2, they need to speak on-topic for all three questions and for a B2.1, they need to speak on-topic for two questions.

Although the current study looks only at the B2-level criteria (scores 3 and 4), test-takers range from Pre-A1 to C2-level and the scale allows raters to mark at these other levels. The B1 criteria (scores 1 and 2) have the same language criteria for high and low B1-levels but different on-topic requirements. A score of C1 (5) has its own criteria with an expectation that all three responses are on-topic. Raters can give a score of C2 (6) for performances above C1 level. The British Council added the C2 level to the scale because raters were reluctant to score a 5 (C1). There was a ceiling effect with an expectation that a 5 (C1) implied a perfect response and therefore very few 6 (C2 scores). This is another interesting area of research on this task type, but out of scope for this project. Any performance below B1 level (A2, A1 or Pre-A1) receives a score of 0.

The design of this Aptis B2-level speaking rating scale with on-topic rating requirements arose from an internal research project to revise the Aptis rating scales (Fairbairn & Dunlea, 2017). Aptis had previously used one rating scale across all tasks but encountered rating ceilings because test-takers could not produce a high enough level of language for high scores on some lower-level tasks.

The methods to develop the new speaking rating scale took a performance, theory and measurement approach (Fulcher et al., 2011). The researchers asked raters to identify which salient features were most important to them when marking. The researchers then matched the salient features to CEFR descriptors and adapted them to meet the needs of the task. The scales were confirmed in a field trial study to moderate the opinions of the people who developed the rating scales. The test changed to task specific rating scales as a result, which means that the scales are only valid for marking a specific task type and comparing performance across tasks is discouraged (Fulcher, 1996, 2003; Upshur & Turner, 1999). The design of the rating scales included the 'performance' of raters, the 'theory' of the CEFR and the 'measurement' in a field trial to provide the evidence to use this rating scale for this specific purpose. The ceiling effect was removed, the scores became more meaningful and topic relevance descriptors were added.

1.2.4 Yes/no binary topic relevance decision in the Aptis Speaking test

This project contributes to the validation of using on-topic rating requirements for scoring test-takers in an extended spoken performance. This section explains the yes/no binary topic relevance decision in the B2-level task in this research project and how it is operationalised.

Aptis uses a yes/no binary topic relevance decision to fine-tune the scoring for the B2-level task. The distinction between the CEFR levels is difficult to determine in an operational speaking test context. Even trickier is the ability to distinguish between high and low CEFR levels within a band (i.e., B2.1 and B2.2). Aptis makes this distinction in the speaking skill by focusing on the ability of a test-taker to sustain the CEFR level on-topic throughout a long turn spoken performance. If a test-taker can

sustain the CEFR level on-topic, the test-taker has achieved the task and is at the high end of the CEFR level (i.e., B2.2). If the test-taker is not able to sustain the CEFR level on-topic but is overall at B2-level, they are at the low end of the CEFR level (i.e., B2.1).

To achieve a score of 4 (B2.2), test-takers need to answer all three prompts on-topic with B2-level language as described in the rating scale (Appendix 4). The reason for a mark of Aptis Score 3 (B2.1) means the response meets the B2 language criteria but is at B2.1 level for one of three reasons:

1. The test-taker does not answer one of the question prompts.

2. The test-taker uses language below B2 level for one of the question prompts.

3. One of the question prompt responses is off topic.

It is important to note that the only difference in the rating scale descriptors for this task between B2.1 and B2.2 is the number of questions the test-taker has answered on topic at B2-level. The rest of the criteria are the same.

Raters train to make a two-part decision in the following order:

Decision 1: CEFR level

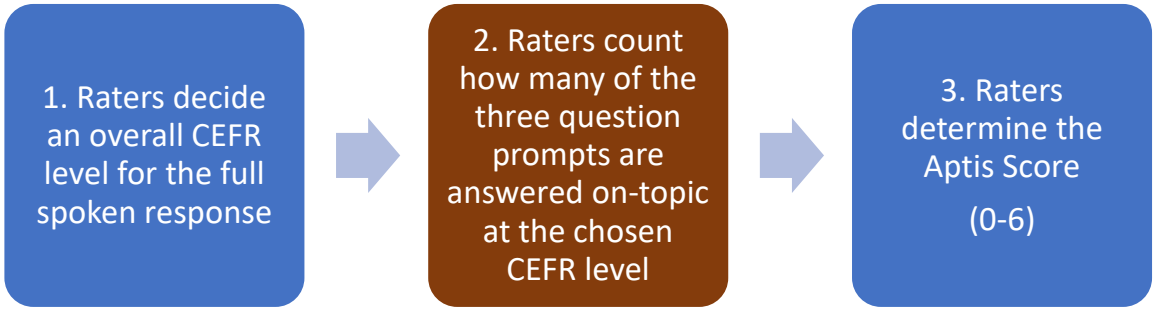
Raters assess the language features (grammar, vocabulary, pronunciation, fluency and cohesion) and decide on an overall CEFR level (Above C1, C1, B2, B1 or below B1). Raters mark holistically and arbitrate between the various CEFR levels of the language features described in the rating scale (Appendix 4) to arrive at an overall CEFR level. The values each rater gives to each trait in the scale may be idiosyncratic and raters may assign the same overall scores to performances but for

different reasons depending on their internal perceptions (Hamp-Lyons, 1995; Iwashita et al., 2008; Knoch et al., 2021). This first decision narrows the possibility of scores based on the CEFR level (i.e., a B2-level performance can only score 3 or 4).

Decision 2: Sustained on-topic response

Next, the raters decide if the test-taker has produced on-topic language at the chosen CEFR level across the three questions. This is a primary trait or atomistic approach to scoring (Goulden, 1992; Huot, 1990a, 1990b; Weigle, 2002) with a yes/no binary decision made by the raters for each of the three question prompts in the task. The rating process can be seen in Figure 1. The second decision in brown is the focus of this thesis.

Figure 1: Aptis speaking test scoring process



This thesis focuses on the internal criteria raters report using to decide if a speaking response is on-topic. Included in the research is how different task prompt types and different task versions from different test forms affect raters' decision-making. I study how the internal criteria that raters reportedly use reflect the criteria in the Aptis rating scale descriptors and if the decisions are common across all speaking test prompts and versions.

This is a mixed methods study so incorporated in this thesis is an exploration of the relationship between the internal criteria raters reportedly use to make their on-topic decisions and the quantitative severity and reliability of on-topic and Aptis scoring. This thesis helps to understand task features, rating scale criteria and rater processes that lead to high or low rater reliability when making the yes/no binary topic relevance decision. Taken together, the current study contributes to knowledge by providing evidence on if using a yes/no binary topic relevance decision for measuring spoken performance is a valid measure of performance.

1.3 Speaking corpus, ethics approval and data protection

To undertake this research, I used the British Council corpus of 9000 Aptis General speaking tests (British Council, 2018). The British Council is the Data Controller for the current study and the data is operational test data used by the Assessment Research Group at the British Council for research purposes. Each test is approximately 8 minutes long. The corpus has been approved by the British Council's Senior Information Governance Advisor (Privacy & Security) for test validation research purposes. This project secured access to use this data and ethics approval to GDPR standards from the British Council and the UCL's Faculty of Education and Society (IOE) (March 2018) (see Appendix 5).

The first ethical matter raised in the ethics approval was the storage of data. The test taker spoken responses from the corpus are stored on the British Council encrypted server and password protected. Speaking files were shared with raters using iSecure encrypted file transfer software and the files have been deleted on iSecure. The questionnaire was issued using the British Council's encrypted questionnaire software and has been deleted. Anonymised test-taker and rater data (scoring and

reasons for on-topic decisions) are stored on the UCL network and the researcher's personal computer. The file linking rater codes to their names are in a separate folder to the folder holding their marks and justifications. Email exchanges with raters are stored on the UCL server.

The second ethical matter raised were how the research would be used in publication. All data is anonymised so that test takers and raters will not be identified. The research has used pseudonyms for the test takers and raters. Test taker speech samples, if necessary, would be transcribed for conferences so that voice recordings are not shared. No ethics or data protection issues were identified by UCL or the British Council in the use of this data.

The third ethical matter was regarding any conflict of interest between my research and my role as a British Council employee. The main purpose of this research for the British Council is for Aptis test validation and to contribute to assessment projects (e.g., automated assessment by helping to determine the importance of a response being on-topic at B2 level). The work will be in the public domain, shared with the British Council and the wider testing community as journal articles and in technical reports on the British Council website. The output from this research and all data will be the IP of the British Council. No conflict of interest was found by UCL or the British Council.

The corpus was collected in 18 countries and questionnaire data on the test-takers were also collected. The scoring was done in the same way as regular Aptis scoring and the same as the scoring in my study. See the pilot studies (Section 3.3) for details on how the test-taker responses were selected for the main research project.

The speaking files are held on a secure British Council encrypted server and anonymised using a file name that identifies the country where the test was taken, the task number (four tasks for each speaking response), the task version, the score for each task, the overall CEFR level for the full test and a unique response number for each task.

Example: Spain_T4_Version 1_4_B2_1489069.mp3

The test-taker speaking response file names were further anonymised to provide even more security and to remove the response information from the file name. I changed the original speaking file name to a response number (e.g., #V1-101). The 'V1' refers to the test version (form) so that the rater uses the correct task prompt.

The speaking responses were sent to the raters using iSecure, the British Council's secure online file transfer platform, along with instructions sent by email (Appendix 6). Raters signed a non-disclosure agreement with the British Council in their employment contract.

2 Literature review

The purpose of this review is to synthesize and consolidate information on how speaking on-topic in language tests has been conceptualised, defined, measured and interpreted in the research literature and by testing organisations. This comprehensive overview and conceptual framing of the on-topic construct definition and operationalisation also explores in more detail the concepts raised in the first chapter.

Section 2.1 provides a discussion on task topics and complexity. I first provide an overview of the history of topic relevance in the literature. I then examine how task topics and complexity can affect test-taker spoken responses and the gap in the research on how task topics and complexity may also affect raters' topic relevance decisions. Section 2.2 investigates the topic relevance construct in spoken performance. Using the Bachman and Palmer (2010) model to frame the discussion, this section uses examples from well-known international language proficiency tests to explore topic relevance scoring approaches. The evidence shows that the construct of topic relevance is not consistent across test rating scales. Section 2.3 considers rater variability when making on-topic decisions. This section includes discussions on rater reliability, the construct relevance of on-topic decisions, research from testing the writing skill and integrated testing, which have a more accomplished topic relevance body of literature rater, which may be useful for understanding topic relevance in speaking, the effects of variability of rater cognition on marking and how test delivery modes could affect raters when scoring topic relevance criteria. Section 2.4 outlines the research questions for the study. The goal of this literature review is an enhancement of our understanding of how rating criteria incorporates topic relevance and how raters manage this part of the rating process.

2.1 Task topics

The first area to investigate in this literature review is how task topics and task complexity affect rater topic relevance decision-making. This section explains how task topics and task complexity can affect test-taker performance and discusses how topics and complexity might also influence topic relevance decision-making by raters. The gap in understanding how task topics and task complexity affect raters is emphasised.

2.1.1 History of topics

Fulcher (2003), Isaacs (2016), Spolsky (1990) and Weir et al. (2013) provide overviews of how the content of spoken performance became increasingly more important throughout the 20th century. Remarkably, until the 1930s speaking assessments in large scale tests in the United States tended to be objective tests with test-takers listening to and identifying correct spoken language. Language testers were sceptical about speaking tests because of perceived subjectivity and reliability issues (Weir et al., 2013, p 27). It was not until the 1930s that *The College Board's English Competence Exam* first introduced test-taker performance rating criteria. Overseas students took this test to study in the United States. The rating criteria focused on how a test-taker spoke and less on the content of what they said. Examiners chose the topics and made three-point decisions (proficient, satisfactory, unsatisfactory) on eight areas of spoken production including fluency, responsiveness, rapidity, articulation, enunciation, command of construction, of connectives, usable vocabulary and the use of idiom; examiners also reported on if the candidate appeared diffident or shy (Spolsky, 1995, p 57). There were no criteria for these eight areas and decision levels and there was likely no training for

examiners. The University of Cambridge Local Examination Syndicate (UCLES) in the UK started testing speaking in 1913 with the Certificate of Proficiency in English (CPE) test, which included reading aloud and conversing on works by English authors (Weir et al., 2013). Rating criteria focused on correct pronunciation, sentence rhythm, intonation and comprehension (Taylor & Galaczi, 2011; Weir et al., 2013, p. 25) and there were no prescribed test settings, for example the amount of time for the speaking test (Spolsky, 1990). There was also a phonetics paper that tested knowledge of how words sound (Weir et al., 2013, p. 262). In the United States and the United Kingdom, there was little if any thought on how topics could enhance or impede spoken response, in particular the impact of unfamiliar topics on test-takers.

The Foreign Services Institute (FSI), Interagency Language RoundTable (ILR) and American Council on the Teaching of Foreign Language (ACTFL) oral interviews in the 1950s and 1960s were the first to include test content as criteria for rating scales (Fulcher, 2003; Higgs & Clifford, 1982). Two approaches to measure spoken language proficiency with respect to topic relevance appeared during this time, which the Common European Framework of Reference (CEFR) at the Council of Europe (2001, 2018) still uses to differentiate levels of performance:

1. As a test-taker becomes more proficient in speaking, the ability to deal with unfamiliar and diverse topics changes (de Jong, 2023). At low proficiency levels, test-takers are only able to talk on topics about their own basic personal circumstances. Lack of topic knowledge and vocabulary in short-term memory in low proficiency level test-takers is one area that affects the ability to deal with unfamiliar topics (see Bachman, 2010, which is further

discussed in the next section). At intermediate levels, test-takers can speak on topics in their field of interest. At the highest-level levels, they can speak about unfamiliar topics because they have the vocabulary in their short-term memory to deal with unfamiliar topics.

2. These scales were the first to identify how concrete and abstract topics differentiate levels of spoken performance. At higher levels, test-takers can respond to test question prompts that include intangible abstract topics such as feelings.

However, it should be noted that the criteria in the CEFR for understanding and handling topics may not be well linked to the quality of spoken performance. As Pallotti (2009) and Leung (2010) point out, it is possible to speak adequately about abstract topics using relatively simple forms of language and vice versa.

Pallotti (2009) defines adequacy as the degree to which a test-taker's performance successfully achieves a task's goals efficiently and views the concept of adequacy as a separate performance indicator from the raters' perceptions of the complexity, accuracy and fluency (CAF) or linguistic characteristics of the spoken response. Test-takers can produce language that is neither complex, accurate nor fluent but achieves the intended task goals. It is possible to achieve a task's goals efficiently and speak on-topic without necessarily developing the topic or knowing anything about the topic. How task achievement (i.e., adequacy or communicative success) is related to topic relevance is under-researched in the literature.

It was during the next two decades in the 1970s and 1980s that there was a shift towards the concept of communicative competence. This new approach linked test tasks to the context of situation, sociolinguistic knowledge and culture (Bachman &

Savignon, 1986; Canale & Swain, 1980; Hymes, 1972) of knowing “what to say and how to say it” (Révész et al., 2016, p. 830). Communicative competence was controversial at first because assessing spoken production became more qualitative and subjective with degrees of compliance (McNamara, 1996; Shohamy, 1983). The models of spoken performance developed by Bachman and Savignon (1986), Canale and Swain (1980) and Hymes (1972) during this time approached topic relevance in the context of communicative competence in distinct ways.

Hymes (1972) developed a theory of communicative competence that integrated linguistic theory with communication and culture. The focus of communication moved away from grammatical perfection outlined by Chomsky (1965) and towards the appropriateness of the spoken production in the context of a situation and culture. Competence, according to Hymes (1972), includes knowing what to talk about and in what manner, and involves attitudes, values and motivations. It is about language use and performance.

The theory included four questions to help determine if an utterance is communicatively competent. Is the spoken response 1) possible; 2) feasible; 3) appropriate; and 4) performed? There are interrelationships amongst the four questions and degrees of compliance. The aim of these four questions is to help understand utterances with respect to the notion of ‘acceptability’. When Aptis raters assess test-taker responses and decide if the response is on-topic or off-topic, could they be, consciously or subconsciously, asking themselves one or more of the four questions developed by Hymes (1972)?

Widdowson (1989) concluded that answering the four questions developed by Hymes (1972) is complicated. According to Widdowson (1989), there are degrees of

acceptability in this model that are difficult to specify and standardise. Moreover, speaking topics and interlocutors affect performance. Awkward sentences are “illegitimate, rule-violations, not possible” (p. 133) but could be acceptable to some people.

Canale (1983) developed a model of communicative competence that included grammatical, sociolinguistic, discourse and strategic competence. There is no separate criteria for topic relevance. Instead, topic relevance is dispersed amongst other criteria. Lexical knowledge, thought to be important for topic relevance, is categorised under grammatical competence. Sociolinguistic competence implies topic relevance for different cultural situations. Discourse competence includes coherence and cohesion. Strategic competence includes linguistic strategies to overcome communication issues, such as using pre-prepared responses that are tangentially topic relevant.

This model is relevant for my current study by highlighting that Aptis raters could be using other parts of language to make their on-topic decisions. For example, the Aptis rating scales include the identification of cohesive devices in the language criteria, but there is no explicit criteria for coherence (see the rating scale in Appendix 4). It is not clear if or how raters decide if a response is coherent. Raters may assess coherence as part of the language criteria, as part of the their on-topic decisions or not at all.

Bachman and Palmer (1996) developed a third model, which defines language ability as consisting of language knowledge and strategic competence and called it ‘communicative language ability’ (CLA). The model also includes the test-takers’ personal attributes, affective schemata, cognitive strategies, topical knowledge and

the characteristics of the tasks. I will discuss this model in more detail with respect to topic knowledge and task characteristics in the next sections.

2.1.2 Test-takers and topics

The importance of task topics in question prompts evolved over time as the impact of topics on test-taker performance became more established. This section consolidates the literature on how task topics can affect test-takers' spoken performance.

It is a long-held view that test-takers find unfamiliar topics difficult to manage. Bachman and Savignon (1986), in a discussion of the ACTFL speaking test, included topic knowledge and task content impacts as test method factors and discuss how the topics in the question prompts can affect individual spoken performance. Robinson (2015) includes the number of elements to address in a spoken response, expected grammar output and topic knowledge as part of task complexity. Test-takers must be able to plan what to say, quickly retrieve the required vocabulary, build correct grammatical structures, pronounce the words comprehensibly, use connected speech intelligibly and monitor spoken performance (de Jong, 2023). This psycholinguistic approach to spoken production dominates speaking assessment in language testing (Roever & Kasper, 2018) and views speaking as an individual cognitive skill, which involves the mental processes mentioned above (Godfroid & Hopp, 2023). If the topic is unfamiliar, these psycholinguistic processes become strained.

Bachman and Palmer (2010) refer to the extent topic knowledge is available to test-takers in their long-term memory as the reason test-takers may struggle to respond to a task prompt. An unfamiliar prompt affects a test-taker's response because it puts

a strain on memory. Processing language, or attention, becomes difficult if there is a strain on memory. Automaticity, or the ability to retrieve information seamlessly from memory, becomes difficult and affects the quality of spoken language (Levelt, 1989; Van Moere, 2012). The cognitive load on the test-taker required to deal with an unfamiliar topic affects the test-taker's ability to achieve the task (Skehan, 1998). Therefore, if a test-taker does not have the background knowledge and vocabulary to answer the questions, they may perform poorly. Test developers consider unfamiliar topics to be a form of test bias in a general proficiency test where all topics should be familiar to all test-takers.

In a speaking test, the topic sets the context, situation and cultural setting of the task, which can help or hinder test-taker performance. Test-takers may be familiar or unfamiliar with the context, for example when responding to speaking question prompts that assume local knowledge, geography or customs. Topic familiarity and content knowledge therefore become aspects of context validity which test developers need to examine. Investigating topics and context validity has been well-studied in test-takers, for example Lumley and O'Sullivan (2005), Seedhouse and Harris (2011) and Khabbazbashi (2021).

Lumley and O'Sullivan (2005) researched how topic knowledge affects test-takers groups differently. These researchers were interested in the interaction between the task topic, the gender of the test-taker and the gender of the perceived rater. Does the task topic affect male or female performance? Do test-takers perform differently if they think they are speaking to a male or female examiner? They used task topics with stereotypical male or female themes and swapped gender roles. The test-takers spoke into a tape recorder to a perceived male or female rater. A group of 30 raters

marked 894 test-taker performances across 14 versions (forms) of the test. Each individual task was included in several versions of the test in order to compare tasks across versions. Using Many-Facet Rasch Measurement to determine bias against or in favour of any specified group, the researchers found small yet significant task effects related to gender. While the test-takers may perform differently based on their gender and the topic of the speaking test, they found that test-takers do not perform significantly differently if the examiner is perceived to be male or female. This research provides evidence that the test-takers' background information about the topic of the speaking test can affect test-taker performance but that the gender of the rater has minimal effects.

Seedhouse and Harris (2011) investigated how low-, medium- and high-level test-takers develop topics in the IELTS test (see Section 2.2.1 for a description of the IELTS test). These researchers also studied if the examiner's interactional style and adherence to the topics affected test-taker performance, and if certain topics were problematic for test-takers. To obtain a high score on the speaking test, test-takers need to understand the question they have been asked, provide an answer to the question, identify the topic inherent in the question and develop the topic (Seedhouse & Harris, 2011, p. 15).

The researchers used conversational analysis to analyse 130 recordings with a variety of proficiency levels and topics. IELTS has two parts (Parts 1 and 3) that are interactional between the examiner and test-taker and one part which is an individual long turn monologue by the test-taker with one minute of preparation (Part 2). Conversational analysis was not fully possible for Part 2.

Part 2 in the IELTS test closely resembles the task in my research (see Sections 1.1.2 and 1.1.3 for a description of the Aptis speaking test and the B2 level monologue speaking task used in this research). In Part 2, test-takers speak on a topic, which is spoken by the examiner and written on a card. The examiner does not normally interact with the test taker. The examiner may ask one or two scripted rounding-off questions related to the topic when the test-taker has finished speaking.

Topic development is expected in the IELTS speaking test in the 'Fluency and coherence' and 'Vocabulary' criteria. At level 9 on the 'Fluency and coherence' scale, which is the highest possible score on this criterion, "Topic development is fully coherent and appropriately extended" and at level 8 "Topic development is coherent, appropriate and relevant" (IELTS, 2023). The 'Fluency and coherence scale' does not mention topics below level 8. However, under 'Lexical resource', there are topic criteria at each level in the scale. The criteria include the ability to use vocabulary to speak about familiar and unfamiliar topics depending on the level, as well as the flexibility and precision of vocabulary chosen for the topics. The operationalisation of topic relevance in the IELTS speaking test is reflected in the vocabulary used by the test-taker and assumes topic development and coherence at the highest levels.

The researchers identified key features of topic development for low- and high-level test-takers in Part 2. High-level test-takers developed topics and sub-topics and had shifts of perspective (e.g., discussing illness from a patient's and doctor's perspective) and tended to construct an identity in their response of a high achieving global individual. Lower scoring test-takers had more instances of topic trouble, which included short monologues, lack of coherence, repetitive vocabulary and lengthy pauses.

Seedhouse and Harris (2011) noted the difficulty in defining what constitutes a topic and the complexity of systematically investigating adherence to topics. Test-takers transition from one topic into another, which can be stepwise or disjunctive changes (Seedhouse and Harris, 2011, pp. 7-8). Stepwise, or tangential, topic transitions occur when there is a change of topic but the flow from one topic to the next topic is linked. Disjunctive topic changes, or boundaried topic shifts, have larger distances between topics than in stepwise transitions.

Seedhouse and Harris (2011) also found evidence that some question prompts were problematic. Some test-takers had difficulty answering questions which required them to imagine a scenario that they had never considered (e.g., Would you like to be in a film?). There was also evidence that test-takers struggled with questions that were too intrusive or intellectually challenging.

Khabbazbashi (2021) also investigated test-takers and topics and found that, similar to Lumley and O'Sullivan (2005), the background knowledge of the topics by individual test-takers can affect performance. Khabbazbashi (2021) used four tests that comprised of IELTS tasks, four raters and a hybrid IELTS and TOEFL iBT rating scale in a mixed methods parallel forms reliability design study to identify if topic knowledge affects test-taker performance (see Sections 2.2.1 and 2.2.2 for descriptions of the IELTS and TOEFL tests and their rating scales). IELTS was chosen because these are independent tasks and can therefore isolate the background knowledge variable. Independent tasks are short question prompts which require test-takers to use their background knowledge to respond (i.e., there is minimal input in the question prompts). The researcher used IELTS language criteria and TOEFL topic development criteria for the marking scales.

The researcher used a panel of expert raters and an equivalence checklist to review a selection of IELTS tasks and chose five tasks for each of the four speaking test versions. IELTS has three tasks in the test (tasks A, B and C) and the test used in the research included two task As, one task B and two task Cs. The only difference between the four test versions were the perceived decisions about familiar and unfamiliar topics. The researcher controlled all other variables, including discarding task versions with unfamiliar grammar or lexis. Two versions were randomly administered to 82 test-takers in an incomplete-connected data collection design with overlapping coverage of 18 different topics. Two common A tasks were allocated to each test-taker to be able to link the data. The researcher elicited spoken performance on 10 topics from each test-taker.

Khabbazbashi's (2021) research used hierarchical regression analysis and Rasch analysis to examine the extent to which background knowledge as a variable can predict spoken performance. The hierarchical regression model isolates and analyses how different variables affect the scoring outcome and enters predictor data into the model in blocks. The first block was the general language proficiency estimates, which were collected using C-tests administered to participants. C-tests are written texts with distortion, for example the removal of words. Test-takers draw on their knowledge of the language to fill in the missing words. The second block was the results from a background knowledge questionnaire. The questionnaire included 80 questions and elicited test-taker self-perceptions regarding the topics, and was collected after the IELTS tasks were delivered. The third block were the results on the test versions. Rasch analysis was used to examine the quality and functioning of these three data collection instruments and as a way to construct linear interval-level measures so that the data could be compared.

The main quantitative finding was that test-taker background knowledge about topics had significant but small effect sizes on overall test-taker scores. General language proficiency accounted for 60.1% of the variation in spoken performance on different topics (Khabbazbashi 2021, p 151). Including background knowledge in the model increased the variation to 63%. The effect size on scoring is small for background knowledge, compared to general language proficiency, showing that raters and/or the rating scales put more emphasis in their scoring on language proficiency than on topic relevance.

Khabbazbashi (2021) was able to separate the test-takers into statistically distinct ability levels with respect to topic development using separation indices and strata estimates from multi-facet Rasch analysis. The findings show that there were 4.7 speaking proficiency levels for the TOEFL iBT topic development rating, which was discriminating but much lower than for the IELTS language rating, which had 12.34 separate levels (Khabbazbashi, 2021, p. 117).

The qualitative findings in Khabbazbashi (2021) showed that some test versions with the same construct and the same difficulty level elicited different language functions. Functions refer to the type of spoken language such as describing, comparing or expressing preferences. Khabbazbashi (2021) controlled for functions. The fact that some tasks elicited different functions provides evidence that the inferences drawn based on the same scores may not be valid. These qualitative findings were inconclusive due to the small rater sample size (four raters) and the focus of the research being on the test-takers.

Khabbazbashi (2021) also found that some topics affected different groups of test-takers. Some high level test-takers used strategies to deal with unfamiliar topics,

while low level test-takers had more performance issues if a topic was unfamiliar. In addition, the sample of test-takers was Iranian and two topics were unfamiliar to this test-taker population; there was a lack of local validity for these two topics.

This finding extends previous research by Robinson (2001), who also found that the topic affects performance if the test-takers are unfamiliar with the topic. Robinson (2001), however, did not quantify findings as thoroughly as Khabbazbashi (2021).

2.1.3 Raters and topics

While test-takers and topics have been comprehensibly studied, there is less research on how topics affect raters marking speaking tests. Hamp-Lyons and Mathias (1994) found that raters might make different scoring decisions when marking writing tests if they perceive the topic to be easy or difficult, but the research does not extend to topic relevant decisions or to the speaking skill. O'Sullivan (2012) noted that a rater can have a reaction, consciously or unconsciously, to a speaking task topic but this research does not go into detail or provide evidence.

Seedhouse and Harris (2011) stated that examiners may ask themselves the following questions (p. 28).

(a) Is the question answered?

(b) Is the topic focus clear or not?

(c) Is there movement from topic to topic and/or sub-topic?

(d) Is such movement stepwise, motivated and flowing, or not?

(e) Is topic development coherent or not?

(f) Do candidates mark topic/sub-topic shift clearly or not?

(g) For what length of time is the topic developed?

However, the researchers did not interview the examiners so it is not clear if examiners do in fact ask themselves these questions. The research conducted by Cai (2015) and Khabbazbashi (2021) are two studies that include a rater persepective about topic relevance and provide some research on what raters perceive about topic relevance.

Cai (2015) researched raters types and rating decisions to identify if raters have distinct preferences for linguistic performance (pronunciation and grammar-vocabulary) or content (topic relevance, richness in content and organisation). The study was in two parts.

In the first part, the researcher classified 126 raters into three types of rater using cluster analysis to a decision-making task. Raters were given only test-taker analytical scores across five criteria (pronunciation, grammar-vocabulary, topic relevance, richness in content and organisation) for 120 hypothetical test-takers and asked to provide a holistic score on a 100-point scale for the overall performance. The raters did not listen to the responses. The analytical scores varied in which criteria had high and low scores.

The researcher used this method to identify which criteria were most important to each rater and separated the raters into three types: form-orientated (linguistic), balanced and content-oriented. Using regression analysis for each rater to obtain weighting patterns, the researcher classified raters using cluster analysis. Form-orientated raters put the most emphasis on pronunciation and content-oriented raters

emphasised topic relevance. Balanced raters divided their decisions between form and content.

The second part of the research by Cai (2015) included 21 raters from the original sample who rated 20-30 second segments of test-taker responses from an independent task in the TEM-4-Oral test used in China. The segments had different form and content patterns. The first response was strong in form patterns including pronunciation and grammar-vocabulary. The second response was strong in content patterns including topic relevance, richness in content and organisation. The third response was strong in both form and content. The fourth response was strong in neither form nor content. The fifth response was a random response used as a reference.

The raters used a 100-point scale in increments of 5 points and marked 10 segments for three criteria: content, pronunciation and grammar-vocabulary. A mean score over 85 was considered excellent and a mean score under 60 was considered a fail. Content included topic relevance, richness in content and organisation. Topic relevance was defined as the degree to which the test taker's response addressed the key themes. Richness in content was the adequacy and amount of detail. Organisation was the logic of topic development.

This second round of research also collected verbal report protocols of which criteria they valued most and why. The data was coded in segments of discourse for analysis. Form-oriented raters prioritised pronunciation and content-oriented raters prioritised topic relevance and content richness, which was similar to the holistic judgement task in the first part of the research. The key finding is that raters have different levels of emphasis that they put on content criteria. Also relevant is that the

researcher found that raters used pronunciation, grammar, vocabulary and fluency to score content.

The researcher had a number of caveats, which affects the validity of this research. One issue is that the initial rater classification, where the raters do not listen to the spoken responses, may not be comparable to live rating decisions. Another caveat is that the raters had just completed standardisation, which may have homogenized the behaviour of the raters. Moreover, the hypothetical scores for each criterion did not correlate with each other in any systematic way. In addition, more than half of the raters had no experience with this test. It was also a small and unbalanced sample size for the verbal protocols, which led to inconclusive findings, and it is not clear how the 21 raters were chosen. The segments were 20-30 seconds long but it is not clear why this length of time was deemed appropriate. It is also not clear if topic choices or test-taker background knowledge affected raters' topic relevance decisions. The scoring criteria was also context specific, so these findings cannot be generalised.

Khabbazbashi (2021), discussed in section 2.1.2 (Test-takers and topics), is another study that included raters, the speaking skill and topics. According to Khabbazbashi (2021), there is a perception by raters that the topic is the most salient feature in a speaking test.

Raters also had different opinions on the role of topics. Some raters were more focused on the linguistic competence and did not put much emphasis on the topic. The topic was only a way to elicit spoken performance in order to mark linguistic features. These raters were not concerned with communicative success or task achievement. Other raters had the opposite view and communicative success or task

achievement was equally or more important than linguistic competence. However, Khabbazbashi (2021) had only four raters in the study and it was not an in-depth qualitative rater decision-making study, so the findings are not significant and we should be cautious with extrapolating the findings from these four raters.

Khabbazbashi (2021) included a test-taker who did not have the background topic knowledge but never-the-less provided an adequate response. This finding exemplifies how a test-taker can still achieve a task's goals efficiently and speak on-topic without necessarily developing the topic or knowing anything about the topic, as also noted by Pallotti (2009).

One explanation for this view is the standardised nature of international general proficiency speaking tests, which constrains the raters on topic choice or how they can direct the speaking interview with respect to topics (e.g., IELTS). In Aptis, which is a semi-direct test delivered by computer prompts, raters have no control over topics. The question prompts are standardised. It is mostly unknown whether this lack of topic choice and ability to direct a speaking interview affects raters' topic relevance decisions.

2.1.4 Task complexity

Task complexity manipulations affect the mental operations or cognitive processes test-takers need to achieve a task such as interpreting input, planning a response and automaticity (Field, 2011; Levelt, 1989; Wigglesworth, 1997). Changing expected verb tenses and targeted functions in test-taker responses, altering the length or the number of elements in question prompts and including different lengths of planning time are examples of ways test developers can change the complexity of a task (Robinson, 2001, 2015). For example, in an interactive speaking task between

an examiner and test-taker, the spoken performance takes place under time pressure and requires quick access to long-term memory (Bygate, 1987). By contrast, if there is planning time, test-takers can plan on-topic content for spoken production which can enhance performance (Field, 2011; Qiu & Bui, 2022).

Perfect and unambiguous spoken communication by test-takers is generally unattainable and test-takers use a range of strategies to avoid difficulties in expressing ideas (Bygate, 1987). Raters cannot probe or clarify these test taking strategies during the Aptis test. Moreover, test-taker performance in a semi-direct test tends to be more formal and less conversation-like because the test-taker is speaking to a computer (McNamara et al., 2002; Stansfield & Kenyon, 1992a).

Sasayama and Norris (2022) manipulated task complexity with an aim to show that increasing cognitive task complexity (or cognitive load) affects test-taker performance. The researchers used three terms to define cognitive task complexity: intrinsic, germane and extraneous. The general requirements expected of the test-taker and how these requirements interact with the test-taker characteristic is the intrinsic complexity (Sasayama, 2016) and is related to the culture in the task question prompts. Germane complexity in a task is the linguistic expectations of the response, for example the expected grammar. Extraneous complexity is the task requirements which distracts test-takers such as overly complicated or ambiguous question prompts.

The research used four picture narration tasks and changed the levels of complexity. The main difference in the tasks was the number of people in each task: one person in task 1, two people in task 2, four people in task 3 and nine people in task 4. The researchers also made other changes to the tasks to change the complexity levels

(e.g., storyline clarity, code complexity and setting/background) and validated the assumed cognitive task complexity differences in the four narrative tasks before starting the study (Sasayama, 2016). Task 1 was considered the least cognitively demanding in terms of intrinsic, germane and extraneous complexity, while task 4 was the most cognitively demanding. Task 2 with two characters was more complex than task 3 with four characters because of higher extraneous complexity.

One hundred and twenty test-takers participated in the study and answered all four tasks by telling stories related to the picture prompts. The tasks were randomly presented to test-takers. Test-takers had 30 seconds to prepare a response to each task and five minutes to respond. There were three raters and all responses were scored by two of the three raters. Test-takers completed a cloze test to measure their English proficiency and a questionnaire on their perceptions of task difficulty and mental effort.

Sasayama and Norris (2022) used a task accomplishment 6-point rating scale to measure content, which includes the implicit notion of topic relevance. Raters assessed completeness, effectiveness (detail and clarity) and creativity (story-telling flair) as criteria on this scale. To receive the highest score (5), there was an expectation of comprehensive coverage of all content in the input and extensive supporting detail. There was also a delivery rating scale which included comprehensibility, language effectiveness and appropriate response structure. On the 6-point task accomplishment rating scale for content, pairs of raters scored either exactly the same or adjacent for 96% of test-taker responses. The researchers also analysed the test-taker responses for linguistic indices including syntactic complexity, accuracy, lexis and fluency.

The researchers found statistically significant differences across mean ratings for the four tasks on task accomplishment. The lowest complexity task (task 1) elicited the highest task accomplishment ratings and the highest complexity task (task 4) had the lowest ratings. Task 2 had slightly lower task accomplishment ratings than task 3.

The number of word types elicited by each task was positively correlated with the number of characters in each story. However, only for task four was this correlation non-overlapping.

Lexis correlated on average the highest with task accomplishment on all tasks. Task 4, the most complex task with the highest number of people in the picture, had the highest Pearson correlations between task accomplishment rating and the linguistic indices with the strongest correlation between task accomplishment and vocabulary (.77). One reason for the strong correlations between vocabulary and the most complex task could be that there is more to discuss in the question prompt.

Task 3 elicited the best test-taker performance in terms of the four linguistic measures because of low extraneous complexity (Sasayama & Norris, 2016), reiterating the importance of good task design.

The researchers divided the test-takers into three groups according to their cloze test score: low, medium and high. They found moderate to strong relationships between test-takers' task performance (accomplishment, syntactic complexity, grammatical accuracy, lexis and fluency) and their cloze test scores. For lexis, the 95% confidence intervals are non-overlapping only for task 4.

The rating criteria in this study is much broader than just deciding if a response is on-topic or off-topic and highlights how the type of topic relevance criteria and expectations of topic development could affect rater decisions.

2.2 Topic relevance construct and rating scales

Speaking on-topic in a spoken performance test needs to be relevant in terms of a topic framework (Brown & Yule, 1983). A topic framework is the construct of how to assess topic relevance in a test and information about the topic framework is included in the rating scale. Topic relevance criteria in a rating scale can range from a narrow construct, where the rater decides if the response is on-topic, to a broader construct where the on-topic decision includes the quality of the spoken response (McNamara, 1996). Rating scales can include the on-topic construct implicitly as part of other language criteria such as vocabulary or coherence, or explicitly in non-linguistic criteria such as fulfilling the goals, expectations or requirements of a task.

Bachman and Palmer (2010) identified three approaches that topic knowledge can play in the construct definition of a test. The first approach is to combine language ability and topic relevance into a merged construct. A rating scale using this approach embeds topic relevance within language criteria. There are no separate scores for topic relevance. The second approach distinguishes language ability and topic relevance as separate constructs. The third approach assesses language ability without including topic relevance in the rating scale. This section discusses the operationalisation of these approaches using well-known English language tests as examples.

2.2.1 IELTS

The IELTS speaking test illustrates the first approach to topic frameworks and uses criterion-referenced rating scales with topic relevance criteria embedded into the language criteria. Criterion-referenced rating scales describe language proficiency or performance at different levels and provide information on what a test-taker can do (Bachman, 1990; Bachman & Palmer, 1996; Bachman & Savignon, 1986; McNamara, 1996; North, 2003). Glaser (1963) was the first to discuss criterion-referenced rating scales and described them as a “continuum of knowledge acquisition ranging from no proficiency at all to perfect proficiency” (p. 519). These types of scales are simplified representations and operationalisations of the theoretical test construct and used by raters to make judgements about language quality (Alderson et al., 1995; Fulcher, 2003; McNamara, 1996; Weigle, 2002).

IELTS has four analytic rating scales: Fluency and coherence, lexical resource, grammar and pronunciation. Raters use these rating scales to judge parts (or traits) of the performance and score each criterion on a 10-point scale (0-9). A test algorithm provides the final score.

There is extensive research on the benefits and drawbacks to scoring analytically, as opposed to holistically (Knoch et al., 2021). Analytical scoring assesses the spoken performance in parts, or traits. Holistic scoring by comparison assesses the speaking performance as a whole performance and the test taker is placed in a category where their performance as a whole best fits based on blended criteria (Hamp-Lyons, 1995; Huot, 1990a, 1990b; Weigle, 2002). In analytic scoring, features of speech are marked independently and can be weighted. Analytic scales can help raters focus on different parts of the spoken performance and avoid superficial

aspects of performance. Goulden (1992), however, warns that the sum of the scores for the parts in analytic scoring does not necessarily equal a valid score for the whole performance. Douglas (1994) found that there can be relative stability to assessing a whole performance and inconsistency when assessing parts.

The ‘fluency and coherence’ descriptors include topic criteria. At the highest level (level 9) on the ‘fluency and coherence’ scale, “Topic development is fully coherent and appropriately extended” and at level 8 “Topic development is coherent, appropriate and relevant” (IELTS, 2023). The ‘fluency and coherence scale’ does not mention topics below level 8.

Under ‘lexical resource’, topics are mentioned at levels 2-8. The criteria include the ability to use vocabulary to speak about familiar and unfamiliar topics depending on the level, as well as the flexibility and precision of vocabulary chosen for the topics. At level 9, topic relevance is assumed. Below level 3, criteria focuses on isolated words and lack of communication. The operationalisation of topic relevance in the IELTS test is reflected in the vocabulary used by the test-taker and assumes topic development and coherence at the highest levels.

The inclusion of topic criteria on the lexical resource rating scale demonstrates the close relationship between vocabulary and speaking on-topic. North (2003) provides an overview of how vocabulary became important in rating scales, for example in the Cambridge UCLES First Certificate in English and the Test of Spoken English. However, North (2003) cautions that using vocabulary to make topic relevance decisions could focus on assessing a test-taker’s memory and may not include the precision, appropriateness or coherence of the response.

North's (2003) view links topic relevance to the coherence of the spoken response, or how the response is organised. However, coherence is larger than topic relevance and implies topic development and pragmatics. Coherence is also a subjective area to assess and depends on other parts of language such as grammar and vocabulary. It is certainly greater than simply speaking on-topic.

IELTS requires test-takers in part 2 of the speaking test to 1) understand the question they have been asked; 2) provide a response to the question; 3) identify the topic inherent in the question; and 4) develop the topic inherent in the question (Seedhouse & Harris, 2011). The first three requirements in this IELTS speaking task are the same for Aptis. Test-takers must understand and answer the questions. The fourth requirement to 'develop the topic inherent in the question' is more ambiguous and not clear in the IELTS descriptors. Topic development is not expected in the Aptis rating scale.

Criteria for the quality of topic development are missing in the IELTS speaking test descriptors, and indeed in most descriptors describing the quality of content (Elder et al., 2017). Without clear criteria, there are effects on the test-taker and the rater. Firstly, as Seedhouse and Harris (2011) point out, test-takers can answer questions without necessarily developing the topics. Khabbazbashi (2021) found through qualitative feedback from test-takers that they are not clear how much to elaborate when developing a topic. Secondly, although there are support materials, IELTS raters may also be unclear about how much test-takers should develop a topic. As Khabbazbashi (2021) noted, raters had mixed views on the strategies test-takers used to deal with unfamiliar topics, such as waffling or going off on a tangent. One

rater in this study noted that there cannot be coherence if the test-taker does not answer the question, linking speaking on-topic with coherence.

In sum, IELTS raters focus on topic relevancy as part of the ‘lexical resource’ and ‘fluency and coherence’ criteria. The inclusion of topic relevance in the ‘lexical resource’ criteria demonstrates the close relationship between vocabulary and speaking on-topic. The inclusion of topic development in the ‘fluency and coherence’ criteria shows an expectation that test-taker responses at higher levels should be well-organised and extends beyond topic relevance to topic development. There is no separate content-oriented or a topic development criterion, which Khabbazbashi (2021, p. 71) documents could be problematic.

2.2.2 TOEFL iBT

The TOEFL iBT speaking test illustrates the second approach to topic frameworks. The topic relevance and language criteria are separate constructs with different rating scales. In the TOEFL iBT test, there are three language descriptors: General description, delivery and language use, and a separate scale for topic development.

Raters arbitrate between their scores in the four analytic scales using 5-point rating scales (0-4) and record one score between 0 and 4 for the full task performance (ETS, 2023). This approach to scoring is different to IELTS analytic scoring where four separate scores are allocated to the response by the raters for the four rating categories and the test developers combine the four scores to determine the final score. In the TOEFL iBT test, the rater is responsible for the final score and uses the analytic scales to guide their decision.

At the highest level (4) in the ‘topic development’ rating scale, the test taker can sustain a well-developed, coherent and clear response sufficient for the task. At levels 2 and 3, the test-taker addresses the task but there is a lack of elaboration. At level 1, the test takers has limited relevant content. At level 0 the response is unrelated to the topic. The similarity between the IELTS and TOEFL iBT scales, with respect to topic relevance, is that coherence is an indicator of high topic development performance in both tests.

There is no mention of vocabulary in the TOEFL iBT ‘topic development’ descriptors. Instead TOEFL iBT includes vocabulary in the ‘language use’ criteria and combines vocabulary with grammar. At the highest level on the ‘language use’ rating scale (level 4), there is a focus on effective use of grammar and vocabulary and reference to automaticity. At level 3, the scale includes coherence as well as range and precision of vocabulary and grammar.

Brown et al. (2005) studied the TOEFL iBT speaking test and found that raters had concerns about appropriate test-taker content. The researchers collected verbal protocols from 10 raters on the content of 40 test-takers’ responses (see Integrated tests section 2.4.2 for a full review of this study). The researchers noted that raters commented on the number and quality of ideas produced by the test-taker but did not agree on which ideas were essential. The raters were more concerned with the fulfilment of the task than the linguistic quality of the spoken performance. This research raises the intersection between topic relevance and task fulfilment. Are they different and do raters internally distinguish between these two constructs when marking tests? Brown et a. (2005, p. 30) includes ‘on topic’ and ‘completeness of response’ as features of task fulfilment. Moreover, this research shows that raters

may use different parts of language to make their topic relevant or task fulfilment decisions, and these decisions could be different for each rater.

Khabbazbashi (2021) used the TOEFL iBT topic development rating scale and showed how this scale captured non-linguistic and non-content-oriented features of test-takers' performance. Separate topic development criteria can usefully isolate features of performance that are separate from language criteria. Khabbazbashi (2021) also found that test-takers received the highest scores for topic development, compared to other criteria. Khabbazbashi used IELTS language rating scales (see section 2.2.1 for the rating categories), and not the TOEFL language rating scales. Moreover, there were only four raters in this study and it was not a study specifically examining rater behaviour, so we should be cautious with this finding. Nonetheless, it does indicate that raters' beliefs about topic relevance could affect scoring.

In conclusion, separating the topic development criteria from the language criteria has benefits of isolating non-linguistic features of spoken performance. Depending on the purpose of the test, this separation could be useful. While tests with an emphasis on the precision of spoken language might not need this separation (e.g., a test for interpreters), testing the ability to develop a coherent argument with clear progression of ideas might find this distinction useful.

2.2.3 Aptis

Aptis also has a separate scale for topic relevance but instead of a criterion rating scale like in the TOEFL iBT, the choice is a binary yes/no decision (see Section 1.2.3 and Appendix 4 for more information on the rating scale and marking process). There is no criteria for topic development or for the quality of the topic relevance. Binary-choice scales are yes/no type decisions and are thought to require less

interpretation by raters and are therefore more reliable (Knoch et al., 2021; Upshur & Turner, 1995).

Upshur and Turner (1995) developed empirically derived, binary-choice, boundary definition (EBB) type scales. EBB scales require the rater to make a series of hierarchical yes/no decisions before arriving at a final score for a performance. Upshur and Turner (1995) selected eight young learner spoken performances at different levels of performance and used a story-retelling task to develop EBB scales. The spoken performance was approximately one minute. Six raters divided the eight performances dichotomously into high or low performances using impressionistic decisions. Impressionistic scoring has no established criteria and how the rater makes their scoring decisions is unique to each rater (Golden, 1994). The raters next reconciled their decisions and developed criterial yes/no questions to separate the responses into agreed high or low categories. The raters then reviewed the high performances and categorised these into three categories, again using impressionistic decisions. They did the same with the low performances. The raters reconciled their rankings and developed criterial yes/no questions to distinguish between six levels of performance. The researchers were able to develop two EBB scales, one for communicative effectiveness of the recitation and the second for grammatical accuracy.

The scoring process using the EBB communicative effectiveness scale developed by Upshur and Turner (1995) included topic relevance. The first decision the rater makes is about the coherence of the test-taker response. If the rater decides the response is incoherent, then the next question is about topic relevance. The next two

questions differentiate if the test-taker has included one, two or three elements of the story.

If the rater's decision is yes to topic relevance, the following decisions are about fluency and vocabulary. Similar to IELTS and TOEFL, coherence is considered a higher-level indicator of performance. The decision to have topic relevance at the lowest level indicates that, according to Upshur and Turner (1995), a response can be on-topic but incoherent.

This thesis focuses on this binary (yes/no) approach to rating topic relevance. The rating is atomistic in that the test developer expects the rater to assess performance using a rating scale that has only two categories (Goulden, 1992). With such a stark approach to making this topic relevance decision without any criteria for quality, there are likely to be borderline and ambiguous on-topic decisions. For example, Levelt (1989) found that degrees of directness in a spoken response can affect the content of spoken production and affect the listener's (rater's) perceptions. Levelt (1989) also notes that test-taker responses might require raters to make inferences about the ability of the test-taker to answer questions. The language choices made by the test-taker, including the topic, its focus and the manner of delivery, can affect raters' perceptions.

We have seen in the IELTS and TOEFL iBT that test developers use other criteria to make topic relevance decisions, most notably coherence, vocabulary and topic development. However, in the Aptis test, there is no language criteria, no description of quality and no number of information units required to make an on-topic decision. This study aims to isolate these topic relevance decisions in the Aptis rater pool to better understand this area of decision-making.

There is research to support and reject marking topic relevance as a binary yes/no decision. Seedhouse and Harris (2011), for example, found evidence that test-takers at the higher end of a scoring scale tend to have fewer instances of on-topic difficulty and more instances of extended turns in which they developed the topics. Test-takers can speak on-topic coherently for the full turn using appropriate lexis and syntax. This provides some evidence that linking speaking on-topic in the Aptis test is indicative of higher-level performance and theoretically sound.

North (2003) notes that using quantitative approaches to assess spoken production can be dangerous. Raters may depend on counting errors or achievements to assess language. It may seem like a suitable way to manage the complexity of spoken language but the outcome could be that raters overly focus on a surface-level and narrow area of performance. However, North (2003) also warns that scales should have concrete and tangible descriptors, and to avoid words such as ‘rarely’ or ‘usually’. There is a contradiction in these two views. There is a risk in using scales which have degrees of performance, but also danger of using binary yes/no decisions. I explore this area and the operationalisation of different types of topic relevance rating scales in the literature review.

A final area to discuss regarding on-topic scoring variability is the impact of delivery mode. In the IELTS test, the examiner can ask follow-up questions or back channel to encourage more on-topic output and stop a test-taker if they speak off-topic or for too long about a topic (Seedhouse & Harris, 2011). The examiner could also negatively affect the test-takers’ performance and support can be inconsistent (Chapelle & Douglas, 2006; Lazaraton, 1996; Stansfield & Kenyon, 1992a, 1992b). Speaking to a computer in the Aptis test, without an interlocutor, means that there is

no opportunity to probe, confirm, clarify, rectify or direct the response and therefore a rater cannot correct speaking off-topic issues during the test. In both the IELTS and Aptis speaking tests, the test-takers and examiners do not choose the topics. For the IELTS test, examiners cannot deviate from the questions and follow-up questions, so there is a limit to how much an IELTS examiner can assist a test-taker struggling to speak on-topic.

2.2.4 Occupational English Test

The final approach is a rating scale that does not have explicit topic relevance criteria. There are two types of tests that might omit topic relevance criteria. The first type of test is where language is the only important element, for example a test for interpreters. The focus is on testing areas such as precision of pronunciation and vocabulary. Speaking on-topic is an expectation which test developers do not need to explicitly mention in the rating scales. The second type of test is where test developers are measuring if test-takers can perform a task effectively. In these types of tests, the question prompts are more than just tools for eliciting linguistic performance (McNamara, 1996; Norris et al., 2002; Sasayama & Norris, 2022). The expectation is that test-takers must speak on-topic to complete the task so test developers do not require topic relevance in the rating scale. These types of tests are for specific purposes, such as testing language in the context of an occupation. They are not general proficiency tests.

The Occupational English Test (OET) is a test that does not have explicit topic relevance criteria. The OET is a test administered to health professionals (e.g., doctors, nurses, dentists and vets) who wish to immigrate and practice their profession in Australia. In the context of a health care environment, the speaking test

measures the ability of the test-taker to achieve a communication task and involves the test-taker performing a profession-specific role with the interlocutor taking on the role of a patient or a relative of a patient (Jacoby & McNamara, 1999; Lumley & McNamara, 1995; McNamara, 1996). Raters score test-takers on a six-point rating scale for four linguistic categories: intelligibility; fluency; appropriateness of language; and resources of grammar and expression. There is no specific topic relevance criteria in the rating scale. The OET is a test for specific occupation purposes and the topics are familiar to the test takers.

The OET rating scale uses the phrase “appropriate register, tone and lexis for the context” for criteria under appropriateness of language. The focus of the rating decision is on how appropriate the test-taker addresses different stakeholders (e.g., patients, relatives or other healthcare workers). Raters train to assess register (using the correct language with different audiences) and tone, as well as the appropriate use of vocabulary required for a medical scenario. This scale is similar to IELTS and TOEFL iBT in using the appropriateness and effectiveness of vocabulary to assess high performance, with a focus on the ability to explain technical matters in lay terms to some audiences.

In sum, this section used the Bachman and Palmer (2010) approach to topic frameworks and four international tests, IELTS, TOEFL iBT, Aptis and OET, to discuss how tests can include topic relevance in test constructs either as part of the language criteria, as separate criteria or not at all. What is notable is that there is a strong link between topic relevance and coherence, vocabulary and topic development.

2.3 Rater variability when making topic relevance decisions

The subjectivity of rating spoken performance is the reason raters may have variable ratings. Alderson (1993) noted that professional judgements by language testers are frequently in conflict. Even after training, raters still exhibit systematic variations in rating decisions (Knoch, 2011; Weigle, 1998). Lumley (2005) explains these variations as the consequence of the struggle raters have of reconciling the features of the spoken response, the wording in the rating scale and the raters' intuitive impression of the response. Wolfe (2006) approaches this phenomenon similarly by discussing how raters have different mental images of the test-taker's production and these images map onto different mental images of the rating scale. Lumley (2005) and Wolfe (2006), both working on the writing skill, identified the key issue as a conflict in arbitration between the test-taker response, the rating criteria and raters' beliefs and habits. Myford and Wolfe (2003) and McNamara (1996) emphasise that raters use personal and professional judgement to match criteria to a spoken performance, which is a complex activity that can be subject to disagreement.

It is probable that the same subjectivity and complexity occurs with topic relevant decisions. Raters may approach their topic relevant decisions in different manners and these decisions might not be easily standardised. This section reviews the literature for rater variability, including reliability, construct relevance, cognitive processes and test delivery modes, drawing on literature from speaking, writing and integrated testing.

2.3.1 Rater reliability

The first rater variability area to discuss is reliability. Rater reliability analysis measures the consistency of scoring (Bachman & Palmer, 1996). Testers use

reliability measurement tools such as FACETS software to determine if raters are consistently making the same types of scoring decisions (Eckes, 2015; Linacre, 2018a, 2018b). This type of analysis can also identify problematic speaking tasks or rating scale criteria. Rater reliability analysis includes intra- and inter-rater reliability (Myford & Wolfe, 2003, 2004; Wind & Peterson, 2018). Intra-rater reliability is how consistent a rater is with their own rating. Do they make the same decisions every time? Inter-rater reliability is how consistent raters are with the decisions of other raters.

Rater reliability is part of a wider validity argument (Bachman & Palmer, 1996; Deygers & Van Gorp, 2015; Messick, 1989; Weir, 2005). If the test scores are not reliable, inferences made about a test-taker's ability may be false (Bejar, 2012; Kane, 2013). For test scores to be informative and useful to stakeholders, any score variance must be directly linked to test-taker ability.

In my research, the main measurement issues I explore are leniency, severity, consistency and bias in raters' topic relevance decisions. Leniency and severity occur when raters systematically assign ratings that are, on average, higher (lenient) or lower (severe) than expected and can take place due to characteristics of the rater, the ambiguity of the rating criteria and/or the rating context (Knoch et al., 2021; Myford & Wolfe, 2003). Consistency is measured by unexpected ratings. Raters could also exhibit systematic bias towards types of question prompts, task versions (task forms) or test-taker responses (Eckes, 2015).

There is an abundance of research showing that rating reliability increases with training, for example Davis (2016), who had 20 raters mark 100 TOEFL iBT speaking responses before training and then three times after training and found an

improvement after training but that rater severity did not change. Raters were more consistent in their rating but did not change their severity (i.e., they were consistently severe or lenient). Lim (2011) and Weigle (1998) had similar findings.

I discuss rater reliability, including multi-facet Rasch analysis, in more detail in the methodology chapter (see Quantitative data analysis procedures, section 3.6.1).

2.3.2 Rater construct relevance

Rating reliability discussed in the previous section is a quantitative concept but the scoring decisions derive from the raters' qualitative perceptions of test-taker responses. Does a rater exhibit consistent decision-making behaviour when making on-topic decisions and is the decision-making behaviour construct relevant? This section discusses the literature regarding variability of rater topic relevance perceptions.

Construct relevance means that raters make decisions and allocate scores as expected and intended by the test developers (Chapelle, 1999). By contrast, construct-irrelevant variance occurs when raters systematically use other criteria not in the rating scale to make their decisions (Messick, 1989). Raters can also over- or under-represent a construct by placing too much or too little emphasis on criteria (McNamara, 1996).

Cai (2015) referred to over- or under-representation of a construct as systematic differences in weighting patterns or unbalanced construct representation and found that there were three types of rater decision-making behaviour. The 'content-oriented rater' placed excessive emphasis on topic relevance. The 'form-oriented rater' overemphasised language. The 'balanced rater' took both topic relevance and

language into consideration somewhat equally (see Cai, 2015 in the 'Raters and topics' section 2.1.3 for more detail on this research). Therefore, regardless of the expectations of the test developers, raters may have internal criteria that puts more or less importance on their topic relevance decisions.

In the Aptis task used in my research, raters are expected to use their on-topic decisions to make their scoring decisions. As discussed in section 1.1.4 (Yes/no binary topic relevance decision in the Aptis test), to achieve a score of 4 (B2.2) test-takers need to answer all three prompts on-topic with B2-level language as described in the rating scale (Appendix 4). Not properly using on-topic decisions (i.e., not counting the number of responses that are on-topic) is an example of construct under- or over-representation; raters might place not enough or too much emphasis on the yes/no binary topic relevance decision.

The importance of understanding how and why raters make their on-topic decisions is because these decisions can affect test-takers' speaking test and overall Aptis test scores and CEFR levels. The B2 speaking task score studied in this thesis is used to determine the score and CEFR level for the full speaking test. The speaking test score and CEFR level is used to calculate the Aptis score and CEFR level. On-topic decisions, therefore, can affect the scores test-takers receive on the B2 task, speaking test and Aptis test. If the on-topic decisions are not used as intended by the test developers, perhaps because of variance in rater internal criteria about the importance of topic relevance, the scoring could be inaccurate.

Because of the importance of the topic relevance decision to scoring, it is crucial to understand how the Aptis topic relevance construct works in practice. Is it appropriate to count the number of responses that are on-topic to make scoring

decisions and are raters able to reliably make these decisions? There is no information on what it means to speak on-topic in the Aptis rating scale, implying that the construct definition may not be fully developed. My thesis aims to start to develop a topic relevance construct definition for the Aptis test and contribute to the theoretical understanding of the topic relevance construct in speaking tests more generally.

There is research which investigates and defines a speaking test construct, with some discussion about rater variability when making topic relevance decisions. For example, Brown (2000) conducted verbal protocols in order to investigate which salient features of spoken performance raters use to make scoring decisions in the IELTS speaking test. Eight raters scored and provided verbal protocols for four recorded test-taker responses in a partially crossed design using eight spoken performances. Raters scored the test-taker responses holistically (i.e., one score) across four criteria: effectiveness of communication in relation to a specified range of topic types; grammatical range and accuracy; the ability to talk at length; and functional range. They were asked to verbally justify how they made their scoring decisions in a retrospective verbal report. Raters then relistened to the responses and participated in a simulated concurrent verbal recall. The recording was stopped so that raters could make comments and identify features of the spoken performance which contributed to their scoring judgement. Raters were provided with verbal and written instructions to guide their protocols and encouraged to consider features in the spoken performances that were not included in the rating scales. Raters varied in the length of their verbal protocols and the number of times they stopped the recording. The data was segmented and coded in an iterative manner.

The researcher categorised three types of comments: explicit and implicit evaluative comments about language features; non-evaluative affective reaction comments that were not included in the rating scales (e.g., test-taker attitude); and interviewer focused comments about the behaviour of the interlocutor who had conducted the speaking test.

Relevant to my thesis was the finding that raters were not always able to distinguish between the sophistication of the content (e.g., quality of ideas) and the sophistication of the response structure (e.g., functional skill or discourse complexity), although the researcher notes that raters might not have been explicit enough in their protocols. However, it is noteworthy that quality of ideas was not included in the band descriptors but some raters included the sophistication of the content to justify their scores.

Orr (2002) investigated raters and construct relevance by collecting ratings and verbal protocols from 32 raters who watched two video recordings of simulated First Certificate of English (FCE) speaking tests. This test has two test-takers in a paired interview in each recording. The purpose was to understand the process raters used to arrive at scoring decisions. The research collected retrospective verbal reports from the raters immediately after watching each video recording. The reports were transcribed, segmented and coded using categories appearing in the data.

The researcher found a variety of rater perceptions that were often contradictory across all criteria (grammar and vocabulary, discourse management, pronunciation and interactive communication). There were differences in severity but also incorrect application of the scale, with evidence that raters lack understanding about the model of language ability on which the rating scale is based. Rating criteria that were

more complex and with more detailed description had more varied patterns. Raters also used criteria that were not in the rating scales with four raters having a sum of references to non-criterion aspects above 70% of total references, including references to content, which was not part of the rating scale. Each rater had a unique interaction of factors which led to the awarding of a score.

Papajohn (2002) examined rater processes when arbitrating between spoken test-taker responses and the rating criteria and found a wide range of approaches. The researcher asked nine raters to design a ‘concept map’ of the decisions they make when marking the SPEAK Test of Spoken English, provided by Educational Testing Service (ETS). A concept map represents relationships among concepts or ideas using charts, tree diagrams, step-by-step processes or flowcharts and can be used to visually represent thinking and understanding. Raters could approach the concept map in whichever way they thought appropriate. The trainer explained concept maps to the raters but did not provide an example.

Papajohn (2002) used a hermeneutic approach to analyse the data using senior raters. This approach uses rational debate to arrive at consensus. The raters first agreed which rating principles were most important by reviewing the rating guidelines and then determined which concept map contained these principles. The assessors agreed that the most important principle was to mark the test holistically, which is defined in this research as overall communication including the amount of effort to understand a spoken response. Other rating principles included salient features of a response, reference to rating criteria, sustainability and internalisation of rating criteria. The purpose was to look for commonalities among good concept

maps and arrive at a consensus regarding the strengths and weaknesses of each concept map.

The research found that some raters had a series of steps, others used Likert scales for different features and others used yes/no paths. Some focused first holistically and then funnelled into the detail. Others started with the detail and funnelled out. Some raters did not mention holistic rating. The researcher highlights that raters had their own internal rating culture and habits.

The concept map activity was part of rater training and of the four raters who had good concept maps, only one passed the training, while two raters with poorly rated concept maps passed the training. The researchers provided two possible reasons for these anomalies: the difficulty in analysing one's own rating process and the struggle to express the process in the form of a concept map. What this research however does identify is that raters have different decision-making approaches and even if the approaches appear to be construct relevant, they may not correlate with scoring.

2.3.3 Writing tests

This section reviews the literature for rater variability in writing tests. Writing tests are similar to speaking tests in that they collect samples of test-taker performance, which are marked by raters. Many of the same speaking topic relevance issues raised in the previous sections are also present in writing tests, with some notable differences.

Milanovic et al. (1996) explored raters' decision-marking behaviour in terms of the approaches they employed while marking Cambridge First Certificate of English (FCE) and Certificate of Proficiency in English (CPE) writing tests, and the elements

raters focused on while marking these tests. The FCE writing test is an intermediate level test with a choice of five topics. The CPE is an advanced test, also with five topics.

The researchers used 40 test-taker writing scripts, 20 from each test. The sample included all topics and all five points on the rating scale. The 16 raters in the study marked the tests in the same order.

The data collection methods included retrospective written reports, introspective verbal reports and group interviews. For the retrospective written reports, raters wrote what they were thinking about while marking and how they reached their final score. There were no prompts or checklists and raters completed a report after marking each writing test. The introspective verbal reports were think-aloud stream of consciousness recordings of raters' thought processes collected while marking two writing tests, which were transcribed and coded. For the 30-minute group interviews, the researchers separated the raters according to their professional and testing background. There were four groups with four raters in each group. These were highly structured interviews focused on marking processes, factors affecting rater decisions and comments on the research methods. The researchers transcribed and coded the group interviews.

The researchers identified three rater perception areas which are relevant to my thesis. Raters perceived that the content of the written response, task realisation and vocabulary accuracy and variety affected their scoring.

Raters had visibly subjective personal perceptions about content and the researchers noted that these perceptions may most influence a test-taker's final score. Content included a wide range of perceptions such as the individuality,

liveliness or dullness of written responses. Raters perceived that content was more important than language unless language was inadequate. Some raters were conscience of their bias towards their preferred type of content and noted that they were cautious when marking for content.

While content in this research refers to the quality of the written response, task realisation refers to the extent to which a test-taker's written response meets the requirements in the question prompts. Task realisation issues were usually associated with scripts that were irrelevant or had poor sequencing, structure at sentence level, linking and ideas development. The researchers found it difficult to determine the weighting raters put on task realisation and noted that it cannot be measured in entirely absolute terms. Some raters split test-takers' final mark between task realisation and linguistic attributes, while other raters considered these two areas closely related, i.e., test-takers need to be competent in writing to achieve the task. Raters also noted the importance of becoming familiar with the task question prompts to understand the task expectations.

The researchers recommended for future studies to use only one test and to use raters experienced in marking the test. They also recommend introspective verbal reports while marking as a robust method. This study also shows the importance of considering the quality of the response (content), the adherence to the question prompts (task realisation) and the accuracy and variety of words (vocabulary).

Kondo-Brown (2002) conducted a study to investigate how rater judgements are biased towards certain types of test-takers and certain criteria when assessing Japanese second language writing. There were 234 written responses in the study using three prompts (78 responses for each prompt). Three raters marked each

response and the data was analysed using FACETS software. The three raters had high interrater correlation coefficients between pairs of raters and were highly self-consistent in their scoring. The rating scale included content, organisation, vocabulary, language use and mechanics categories. For the ‘content’ category, key criteria included knowledge, substance and topic development.

Despite high overall interrater correlation coefficients, the raters scored certain criteria and test-takers more leniently or harshly, and each rater’s bias pattern was different. Content, vocabulary and mechanics had significant bias between the three raters. The highest percentage of significant bias was found among the candidates whose ability was extremely high or low. The researchers equally weighted the marking categories in the final score and noted that changing the weight of the content category could affect final scores.

The research also found that bias occurred with one of the question prompts. The researchers speculated that different topics in the question prompts could have contributed to this bias. However, as this was a quantitative study and without any qualitative data, this finding cannot be substantiated. The researchers recommend follow-up research to include collecting raters’ decision-making processes while rating.

Lumley (2005) is another writing researcher who included content in the investigation. Four raters took part in the study and marked 48 writing scripts in two parts, 24 scripts as close to normal rating conditions and 24 scripts with think aloud protocols. Data analysis included correlations and Many-Facet Rasch Measurement. Raters also attended an interview and the data was transcribed, segmented and coded. In the rating scale used in this study, ‘task fulfilment and appropriacy’

includes relevance, appropriacy of content, ideas, meaning and vocabulary as sub-categories. The tasks used in this research elicited everyday general proficiency responses (e.g., giving information or explanations and discussing an issue). Lumley (2005) found that raters can become conflicted between rating criteria that describe the quality of the content and criteria on fulfilling the task requirements.

In another research paper on the writing skill that includes topic relevance, Kim (2011) found that raters had different perceptions of the criteria ‘This essay answers the question’. Kim included 35 criteria in their checklist-type rating scale and a ‘yes’ response was given by the rater if they thought that a test-taker generally met the criteria of the descriptor; otherwise, it was a ‘no’. The researcher trained the raters to understand the word ‘generally’ to mean that they must not feel distracted and must be able to comprehend the performance for each specific criterion. Kim (2011) grouped the 35 questions into writing skills (content fulfillment (sic); organizational effectiveness; grammatical knowledge; vocabulary use; and mechanics). The first question (This essay answers the question) was only allocated to ‘content fulfillment’ [sic]. However, there were 10 other questions on the checklist that were allocated to ‘content fulfilment’. Kim (2011) viewed content fulfilment as a much wider construct than just writing on-topic and answering the question. There is an expectation of topic development.

Baker (2012) studied the relationship between decision-making behaviours and rating writing tests and found that the choice of task type, task topic, and task difficulty may favour raters with specific decision-making behaviours. The exploratory mixed methods research used a questionnaire on raters’ self-perceived decision-making style, write-aloud protocols, frequency at which raters defer making scoring

decisions and evidence of underusing score levels. The task used in the study was an integrated reading and writing task with a 4-point rating scale. There were six raters in the study and 54 test-taker responses.

The write-aloud protocols were segmented by dependent and independent clause and categorised by five decision-making behaviours: rational, intuitive, dependent, avoidant and spontaneous. There were 26 test-taker responses with four decision-making behaviour categories identified and three responses with all five.

The researcher encouraged raters to mark two scores if they were unsure of their scoring decision in order to identify avoidant and dependent raters. The decision to give two scores is evidence of not making a decision (avoidant) and would need to be adjudicated by senior raters (dependent). One of the raters, identified as spontaneous, had a high rate of double scores (10 of 53 test-taker responses). This finding was explained as possibly because providing two scores is faster than making a scoring decision.

The research used central tendency scoring (avoiding extremes) to identify avoidant raters. The researchers reasoned that lenient raters, who avoid giving low scores, also self-perceive themselves as avoidant raters (i.e., raters who avoid giving a failing grade). However, the FACETS data, which identifies rater leniency, did not support this assumption.

The research reviewed nine test-taker responses which had similar write-aloud patterns by raters. These test-taker responses may be idiosyncratic in some way, rather than represent a rater's pattern of decision-making behaviour. Two test-taker responses with more than one avoidant comment had characteristics of level 1 and 2 and was difficult to score. Three responses with rational write-aloud patterns were

related to word count. Four responses with an intuitive comment, followed by a rational comment found no explanation. In addition, the researcher highlights that not all raters had the same write-aloud patterns.

Baker (2012) found discrepancies when triangulating the data from the different sources. Self-reporting bias in the questionnaire and the 4-point rating scale are possible reasons for the discrepancies. Statistical significance between decision-making preferences and rating accuracy was not proven.

Eckes (2012) also studied writing tests and rater decision-making preferences. The research included 33 raters and used TestDaF, a test of German as a foreign language. There were four anchor responses which all raters scored and the raters marked between 20 and 186 responses each in a partially crossed design. There were three rating categories: overall impression, task realisation and linguistic realisation and 4-point rating scales for each criterion. Within each of the three rating categories there were three descriptors for nine categories in total. The research also collected rater types, or raters' preference: Syntax, structure and fluency.

The scoring data was compared to rater types in a matrix. Eckes (2012) hypothesised that cognitive rater types who had a preference for a rating category, would have similar rating severity and bias in operational rating.

Eckes (2012) found that raters place high or low importance on features in the test-taker responses, and the importance of the features to the rater were associated with the severity of the ratings. Less important features were associated with more severe rating.

A study by Kuiken and Vedder (2016) investigated judgements on functional adequacy in Dutch and Italian writing tests. The researchers define functional adequacy in terms of successful task fulfilment. As part of this research, non-expert raters were asked to quantify and classify the number of pieces of topic relevant information in test-takers' response using a 6-point scale. The scale ranged from 'insufficient' to 'extremely adequate' number of relevant information units. Information units could be ideas or concepts which are countable. The research also investigated task requirements, comprehensibility and coherence/cohesion. There were 71 test-takers with A2-B1 Dutch or Italian level proficiency, as well as 35 test-takers with Dutch or Italian as their L1.

Kuiken and Vedder (2016) separated content criteria from task requirement criteria. While content criteria asked raters to count the pieces of topic relevant information, task criteria assessed specific instructions and requirements of the task such as genre and register. While it has been seen in this literature review that content and task requirements can overlap and be used interchangeably, raters in this study were able to distinguish between the two scales.

Whereas Kuiken and Vedder (2016) defined 'content' as the number and relevance of information units, which can be quantified, for Kim (2011) content fulfillment (sic) "assesses the degree to which a writer satisfactorily addresses a given topic. A writer who shows strength in this area generally demonstrates an excellent understanding of the topic by presenting clear and substantial arguments supported by specific examples" (p. 540). According to Kim (2011), 'content' includes a qualitative decision of degree of compliance and is more comprehensive than simply writing on-topic. Neither of these studies were studying only content relevance but their contrasting

approaches and views on what it means to ‘answer a question’ provide more evidence that the construct of topic relevance is not consistent across rating scales.

This finding is different to that of Lumley (2005) who found that raters could not distinguish between content and task criteria. There could be numerous possible reasons for the difference of findings but likely it is related to the rating scale. Kuiken and Vedder (2016) used a rating scale that had separate descriptors for content and task criteria. Moreover, the explicit criteria to count information units that were adequate and relevant steered raters towards content, while the task requirements focused on areas such as genre and register. Task requirements also counted how many questions were answered, which is similar to Aptis, although it does not explicitly say the questions were answered on-topic. Lumley (2005), by contrast, used a scale that does not make content and task criteria so explicit and essentially includes them both in one category as ‘task fulfilment and appropriacy’.

2.3.4 Integrated tests

There is limited research for topic relevance decisions by raters for independent spoken performance but an area of language testing which has a body of literature is in the field of integrated testing. Integrated speaking tasks combine listening or reading input with spoken or written production (Luoma, 2004). By comparison, tasks with minimal input are termed independent tasks. Most speaking test tasks have some integration, even if the question prompts are only one sentence to read. However, in an integrated speaking task, test-takers use the extensive listening and/or reading input to develop their spoken response and are not required to use their background knowledge to answer the questions. Although some of this literature review on integrated tasks focuses on test-taker performance (e.g., Frost et

al., 2012, 2020, 2021 and Huang et al., 2018), some researchers have found that raters consider the topic relevance criteria more deeply in integrated tasks, for example Brown et al. (2005) and Lee (2006).

Brown et al. (2005) conducted an exploratory study into rater orientations on three integrated speaking task types (one reading-to-speaking and two listening-to-speaking) and two independent speaking task types in the TOEFL iBT. The focus of the study was to understand what raters find salient when marking spoken performances, how this relates to different levels of spoken proficiency and how the task types affect these decisions. No criteria or scales were provided to raters in order to investigate their unguided perceptions. They were provided only with the targeted functions and discourse features as well as the reading and listening input text, speaking prompts and task specifications. Ten raters marked 200 spoken performances, eight responses at each of the five performance levels across the five tasks. Raters also provided verbal reports for one performance at each of the five levels on each task for a total of 25 verbal reports. The sample for the verbal reports was randomly allocated and based on the original scores. The data collection procedures for the verbal reports were the same as Brown (2000) discussed in the previous section: retrospective and follow-up concurrent.

Subcategories identified for content included task fulfilment, ideas and framing. Task fulfilment included the extent to which responses were on-topic or addressed the question. Ideas were the amount of speech produced by test-takers in terms of number of ideas and sufficiency for the task. Raters commented on ideas in terms of the functional demands of the task, sophistication of the spoken response, relevance, logic, clarity of arguments and organisation of ideas. Accuracy of the

content was reported by raters but only for the listening-to-speaking integrated task. Framing referred to the structure of the responses, for example including an introduction or conclusion, but these comments were relatively few in number.

Raters agreed that content was an important focus likely due to the strong nature of the tasks. A weak speaking task focuses only on the linguistic quality of the spoken performance (for example, a speaking test for interpreters) and will have fewer content perceptions by raters compared to a strong task with a focus on topic development or task fulfilment (McNamara, 1996). In the Brown et al. (2005) research, the tasks were on the stronger end of this continuum.

A major finding of this study was the general agreement by raters on what content should be included in a spoken response but there were ambiguities and disagreements about what constitutes key information or important detail. There were more content comments for the listening-to-speaking tasks probably due to test-takers' responses being dependent on their comprehension and ability to remember listening input.

The research also found that the type of task affected the rater verbal reports. There was a more restricted range of expected functions for the independent tasks, with more of a focus on the relevance and sophistication of the spoken production. For the integrated tasks, the expected functions were more complex and there was a wider focus on the input.

Lee (2006) found that content, topic development and task fulfilment rating criteria can affect task generalisability. The researcher used generalisability theory to investigate the relative effects of different task types, number of tasks, number of times the responses are rated (i.e., single or double marking) and raters on test-

takers' speaking score reliability. Generalisability theory analyses measurement error and score reliability where there is more than one facet (In'nami & Koizumi, 2016; Knoch et al., 2021; Lynch & McNamara, 1998). The method quantifies each source of variation (e.g., test-takers, raters, task versions, task types, rating scale type), or variance components, from the total rating variability. Researchers can then study the relative effects on measurement error of different combinations of facets by, for example, removing some raters, ratings or tasks. This method allows a researcher to investigate independent and interactive effects of different combinations of facets.

In the study, 261 test-takers completed 11 speaking tasks: four listening-to-speaking, two reading-to-speaking and five independent speaking tasks. Twenty-two raters participated and each response was double marked using holistic rating scales (Lee, 2005). The holistic rating scale used for the integrated tasks included: ideas and detail; fluency; intelligibility; and vocabulary/grammar range and appropriacy. The holistic rating scale for the independent tasks also included the language criteria (fluency, intelligibility and vocabulary/grammar range and appropriacy). However, instead of ideas and detail, the criterion was topic development. Raters marked all task types but were nested within test-takers in a partially crossed design.

One of the main findings was that tasks were a larger source of error variance than the raters. The test-takers were likely rank-ordered consistently across tasks if only the language-related components (fluency; intelligibility; vocabulary/grammar range and appropriacy) were rated. However, when ideas and detail (integrated tasks) and topic development (independent tasks) ratings were included, rank order changed, suggesting different perceptions by raters for these criteria. Lee (2006) also suggests

that task generalisability is more achievable for independent tasks than for integrated tasks.

Frost et al. (2012, 2020, 2021) focus on test-taker performance rather than rater perceptions but the literature includes useful investigation into how content affects test-taker performance and if content should be criteria in rating scales.

Frost et al. (2012) investigated the validity of an integrated listening-to-speaking task using a discourse-based analysis of test-takers' spoken performance. The purpose of the research was to determine if differences in the quality of spoken performances could be seen across proficiency levels and if task fulfilment should be a category on the rating scale. The research used 50 test-taker responses, who were assigned one of two tasks on the Oxford English language test, and focused on the task fulfilment rating scale which has five categories. The top category (5) includes criteria on effectively summarising the input and skilfully distinguishing between the main and supporting ideas. The test-takers were assigned to high, medium and low groups according to their average overall speaking scores by two raters. The test-taker responses were transcribed and coded using three categories: key points, schematic structure and reproduction of input material units.

Despite the small sample and a circularity issue of using the overall speaking score in order to group test-takers, the quantity of spoken production increased with proficiency, which was a similar finding to Brown et al. (2005). However, Frost et al. (2012) found that test-takers tended to repeat the input material and the quality of spoken performance did not significantly change. Integrated tests are often used for academic purposes (i.e., how well students can integrate reading or listening content into an academic spoken performance) and this research identified that test scores

may insufficiently distinguish test-takers with the ability to produce quality academic spoken output.

In a follow-up study, Frost et al. (2020) investigated relationships between the input material, task demands and test-takers' spoken responses by using semantic profiles in an aim to overcome the quality issue previously identified. Maton's (2013) concept of semantic 'wave' profiles suggests that input text should be examined as a whole, rather than at the word or clause level, and is used as a technique to understand the movement between abstract and concrete concepts in listening or reading input text. The reason for using this method to analyse input text is that semantic profiles could affect the quality of spoken performance.

Semantic profiles are developed by analysing the semantic density and gravity of text. Semantic density considers the range of meaning of the text with weaker semantic density having fewer meanings and thus being more concrete. In a speaking test, a question prompt with few possible answers (e.g., What is your name?) has weak semantic density, while a question prompt with a range of possible responses (e.g., What is life?) is more abstract and has strong semantic density. Semantic gravity weakens as speaking test input becomes more abstract with less dependency on context. The Aptis task used in my research has three questions and the semantic gravity is meant to weaken as the prompts move from concrete and weak to abstract and strong semantic density.

The focus of the Frost et al. (2020) semantic profile research was on task 4 in the TOEFL iBT test which includes listening and reading input. Two versions of the task were used and there were 60 test-taker responses per version using stratified random sampling to ensure a spread of proficiency scores and a balance across

genders. The researchers chose 20 performances at high, medium and low overall speaking test scores and used the same discourse analytic measurement method developed in the previous research. The listening and reading input and the test-taker responses were segmented into idea units and coded in the same manner as Frost et al. (2012). The researchers again note the circularity issue between linking overall speaking test scores with the task scores.

Frost et al. (2020) developed semantic profiles by qualitatively describing the input, or stimulus material, and documenting changes in sematic gravity across the two input texts. The semantic profiles of the texts were then compared to the test-takers' spoken performance. The researchers focused on the semantic profiles of input text which only high-proficiency test-takers could reproduce.

Findings from the Frost et al. (2020) research show similar findings from the previous Frost et al. (2012) research; high-proficiency test-takers, on average, reproduce more of the input text. However, there were differences in the concreteness and abstractness of the tasks and how they were presented did affect test-taker performance. This finding supports Lee's (2006) finding that replicability of tasks in an integrated speaking test might be difficult to achieve.

Frost et al. (2021) next examined the strategies used by test-takers in an integrated speaking test. There were 120 spoken performances and 38 verbal reports by test-takers. The researchers used the same coding structure as Frost et al. (2012, 2020) to analyse the test performances and again encountered circularity issues. The researchers used the verbal reports to identify the strategic behaviours test-takers engaged to comprehend and use the input material. The verbal reports were video and audio recorded and collected in their L1 (first language) for low level test-takers.

Findings showed that high proficiency test-takers had better summarising and synthesising strategies because of better note taking, but all test takers were weak in this area with single ideas dominating the output.

The Frost et al. (2012, 2020, 2021) research raises the importance of carefully considering the content of input and the expectations of spoken responses. The researchers suggest that it is appropriate to consider the accuracy of content as part of the construct of speaking ability in a listening-to-speaking task.

Huang et al. (2018) investigated the relationship between topical knowledge by test-takers performance on independent and integrated speaking tasks. The researchers used four topical knowledge tests, two independent tasks and two integrated tasks.

For the main study, 352 participants completed four topical knowledge tests and one combination of the speaking tasks. The topical knowledge tests collected data on if test-takers expected lists of ideas to belong with the topics in the independent and integrated tasks. The integrated tasks were reading- and listening-into-speaking TOEFL iBT practice tasks with similar input characteristics such as number of words in the reading text, listening duration, amount of information and readability level. The independent tasks had short listening prompts, which were developed by the researchers and had the same topics as the integrated tasks. Four tasks (two independent and two integrated) were administered in a counterbalanced manner to avoid topic effects. Test-takers were split into two groups and responded to one independent and one integrated task with different topics. This method helped to increase generalisability for the research. Two additional tasks (one independent and one integrated) were used as practice tests to familiarise test-takers with the format

of the tests. The topical knowledge tests data were collected one week before the speaking test data to avoid test-takers remembering key vocabulary and topics.

The four-point rating scales were different for the independent and integrated tasks with respect to topic development. For the independent tasks, the criteria included coherence, content relevance, idea elaboration, idea progression and relationships between ideas. For the integrated tasks, the criteria evaluated the completeness, relevance and accuracy of the spoken performance. The researchers also collected scores on four-point rating scales for delivery (pacing, intonation, pronunciation, and stress) and language use (grammar use, vocabulary choice, and idea cohesion). Delivery and language use rating scales were the same for independent and integrated tasks.

The data analysis included path analysis, which is a type of regression analysis used to estimate the magnitude and significance of relationships between variables. The four variables in this research were: independent task topic knowledge; integrated task topic knowledge; independent task performance; and integrated task performance.

The researchers found that topic knowledge had a substantial positive impact on speaking test performance on both independent and integrated tasks. Moreover, topic knowledge for the independent task increased in proportion to the increase in topic knowledge for the integrated task and vice versa. There are claims that providing listening and reading input overcomes speaking topic issues for test-takers, but Huang et al. (2018) show that these claims might be unsubstantiated. This research provides empirical evidence that topics can be a significant source of

construct irrelevance variance and thus provide an unfair advantage to test-takers with prior topic knowledge in both independent and integrated tasks.

Inoue et al. (2021) investigated the effects of two different planning time conditions (20 and 90 seconds) in a listening-to-speaking TOEFL iBT test. Seventy test-takers at two proficiency levels (high and low) participated in the study. The focus of this research was on the test-taker performance and if the planning time change affects scoring and the quality of the performance. Test-takers were assigned to one of four versions of the task: the two different planning times and two different versions of tasks. Raters marked the test-taker responses holistically and second raters marked the responses again analytically (pronunciation, pace, language use and topic development). Test-takers completed a self-reported questionnaire on their cognitive processes and strategies and metacognitive strategies. Sixteen test-takers (eight from the low and eight from the high proficiency groups) also completed a stimulated recall. The focus of the stimulated recall was on the test-takers' note-taking behaviour during the listening and planning time and their spoken performance. The researchers did not find any significant findings for test-taker performance for the two planning times.

2.3.5 Rater cognition

As already discussed, there can be systematic rater reliability differences due to raters' beliefs and values about what is salient in a spoken performance but it is also possible that rater variability may be due to differences in psycholinguistic cognitive abilities, such as phonological short-term memory and attention control. Personality characteristics such as self-confidence, levels of anxiety and interest can also feed

into a rater's decision-making cognitive style (Myford & Wolfe, 2003). A rater's level of cognition and their cognitive style may affect scoring.

Raters need to use psycholinguistic cognitive abilities such as phonological short-term memory and attention control to mark tests (Isaacs & Trofimovich, 2011). Phonological short-term memory is the auditory working memory capacity and refers to the ability to retain spoken sequences temporarily in a short-term memory store. Attention control is the capacity to allocate attention efficiently among different aspects of language requirements in the rating scale, different cognitive processing tasks and between different sets of linguistic relationships. Raters must be able to listen to the test-taker's performance with enough working memory and attention to be able to rate the sample of speech.

Isaacs and Trofimovich (2011) document the research linking phonological memory, attention control and speech perception. They hypothesise that there could be a relationship between rating judgements on spoken performance and levels of phonological memory and attention control. Raters with a larger phonological memory and superior attention control could find it easier to rate spoken performance. However, they note that the exact nature of this relationship could go in different directions. Raters with efficient phonological memory and attention control could find spoken reponses to be more comprehensible, more fluent and less accented, and therefore rate higher than those with less efficiency, or be overly sensitive to errors and rate lower.

To investigate this hypothesis, Isaacs and Trofomiovich (2011) investigated whether individual differences in raters' phonological memory and attention control influenced raters' judgements on dimensions of accentedness, comprehensibility and fluency.

There were 40 speech samples and 60 raters in the study. The researchers used cognitive tasks to measure raters' phonological memory and attention control and then collected ratings from the sample of 60 spoken responses. The analysis compared the cognitive variables and rating and included Pearson correlations, t-tests, interrater reliability measures and distribution analysis. An important finding was that differences in raters' phonological memory and attention control did not play a strong role in rating, although some limitations of the phonological memory task is noted. Moreover, the spoken responses were not cognitively demanding and did not require efficient attention control, which was raised as a further limitation.

Khang (2023) approached rater cognition from a different perspective and explored raters' language aptitude as a cognitive reason for rating differences. The hypothesis was that there is a link between raters' language learning aptitude and their ability to rate test-takers' spoken performance. There were thirty-four raters with English as a first language. They rated sixty-three spoken responses from Chinese learners of English plus two L1 speech samples. The task was an eight-frame picture narrative and raters marked a twenty-second excerpt from the middle of the recordings. The raters used a 9-point scale and marked for comprehensibility, foreign accentedness and fluency. The raters also completed an adaptation of the Modern Language Aptitude Test (MLAT) to determine their cognitive and perceptual abilities to learn new languages. The researcher found that language aptitude had a significant relationship with the ratings.

The relationship between the findings in these research projects and speaking on-topic is not certain. It is not clear if rater cognitive abilities such as phonological control, attention and language aptitude affect raters' ability to rate topic relevance.

For a speaking test to be reliable, a rater's decision-making behaviour must be consistent with the construct of the test and, for some raters, a switch away from their preferred decision-making behaviour may cause additional cognitive load (Bejar, 2012; Cai, 2015). Raters can have different approaches to rating and can place different emphasis on topic relevance criteria. Moreover, different grammatical and phonological organisation of words and choice of words by test-takers can change the meaning of a spoken response (Levelt, 1989) and affect rater perceptions. These different spoken response choices as well as differences in situational knowledge (e.g., local geography unknown to the rater) could affect how raters perceive a response and could affect scoring.

2.4 Research questions

The review of the literature has identified gaps in how and why raters make topic relevant decisions. We have a good understanding of how task topics affect test-takers but we are not clear on how raters interact with task topics.

There are different approaches to marking topic relevance as exemplified by the four rating scales discussed in this chapter. We see that coherence, vocabulary and topic development are criteria that emerge as important for topic relevance decisions but that there are different manners to how these criteria are integrated (or not) into rating scales.

I also delved into the research on rater variability, separating the quantitative research of scoring with the qualitative research on rater construct relevance and cognition, and reporting on literature from other types of tests such as writing and integrated testing.

From this literature review, the following research questions were developed for the main study:

1. What internal criteria do raters report using to decide if a speaking response is on-topic?

    1.1. What generic criteria do raters report using that are common across all speaking test prompts and versions used in this research?

    1.2. How do B2-level Aptis speaking task prompt types affect raters' decision-making?

    1.3. How do the task versions from different forms of the test affect raters' decision-making?

    1.4. How do the internal criteria that raters reportedly use reflect the criteria in the rating scale descriptors?

2. What is the relationship between the internal criteria raters reportedly use to make their on-topic decisions and raters' Aptis scoring and on-topic decisions with respect to rating severity, reliability, consistency and construct relevance?

3 Methodology

This chapter presents both the practical aspects of the collection of data and analysis, as well as the theoretical concepts required to justify and explain the analytical decisions in the investigation. Section 3.1 presents the research design. The participants in the research and the procedures to recruit the participants are in section 3.2. The informed consent are in section 3.3. The pilot studies that informed the research methods for the main study are in section 3.4. The data collection procedures and instruments for the main study are in section 3.5. The quantitative and qualitative data analysis methods are in section 3.6.

3.1 Research design

The study design supports the major aim of the research: to understand how and why raters make on-topic or off-topic decisions. The investigation used a fixed mixed methods convergent primary core design as explained by Creswell and Plano Clark (2018). I collected the core quantitative and qualitative data at the same time and from all 30 participants in the study. They were different and equally important data sets. I analysed the data sets separately and independently and then merged the quantitative and qualitative data into a final interpretation and triangulation. The intent for using this design was to bring together and compare the objective on-topic decisions (QUAN) and Aptis scoring (QUAN) by raters with the subjective reasons for on-topic decisions (QUAL) for a complete understanding of the research problem.

The outcome that I expected to attain by mixing the quantitative and qualitative data was a comparison of complementary data on the topic relevance decisions and a triangulation of results. Using mixed method designs is a way to measure and

observe overlapping and different facets of the phenomenon under investigation and understanding complex relationships (Bazeley, 2009; Greene et al., 1989). This approach can also help to improve the validity of the research through convergence, divergence and corroboration for a more complete response to the research questions (Dörnyei, 2007).

I used inductive, abductive and pragmatic research approaches. Inductive analysis looks for patterns in the data without any framework to guide the analysis (Teddlie & Tashakkori, 2009). Abductive analysis looks at the findings and considers reasons for the results (Johnson & Onwuegbuzie, 2004). As this was an under-researched area without an existing framework to follow, inductive analysis was the main approach in the qualitative analysis. I used abduction to try to plausibly explain the reasons for my findings in the mixed methods interpretation and triangulation. I used a pragmatism paradigm to conduct my research.

Pragmatic approaches use multiple methods with different strengths and weaknesses which elicit different types of perceptions and decisions that specifically answer the research questions (Creswell & Plano Clark, 2018; Johnson & Onwuegbuzie, 2004). The strength of the quantitative data collection and analysis methods for the Aptis scoring and on-topic decisions allowed me to understand rater severity, leniency, reliability and bias when making topic relevant decisions and the impact of the on-topic decisions on Aptis scoring. One weakness of a quantitative only study was that I was not able to understand why raters had made their on-topic decisions. The qualitative data collection and analysis of the raters' perceptions on how and why they made their on-topic decisions provided me with an understanding of their severity, leniency, reliability and construct relevance in their reported

perceptions. Combining the quantitative and qualitative together in a mixed methods interpretation allowed me to fully understand the phenomenon and be able to answer my research questions.

The convergent design was conducted in three parts. The first part collected the data and was in three phases. The first phase collected Aptis scores (QUAN), on-topic decisions (QUAN) and introspective reports on reasons for their on-topic decisions (QUAL). There were 30 raters and 45 recorded speaking responses chosen from the pilot studies (see section 3.3 for the pilot studies). Each speaking response addressed three question prompts for a total of 135 on-topic decisions and reasons for decisions. There were 45 Aptis scores.

The second phase in the first part was an analysis of the data for each rater with the intent of developing semi-structured follow-up email exchange questions.

The third phase in the first part of the data collection were the email exchanges. These were conducted within a week after receiving the core data. The purpose of the email exchanges was to clarify decisions and perceptions and collect additional qualitative data with two follow-up questions which were common to all raters (qual).

The second part of the research was the quantitative and qualitative analysis of the data. The quantitative and qualitative strands were equal and kept separate during the analysis using techniques which answered the research questions (Creswell & Plano Clark, 2018). Teddlie & Tashakkori (2009) emphasise the independent nature and separate analysis techniques of this type of parallel or concurrent mixed design.

Part 3 integrated the quantitative and qualitative data into two mixed methods analyses. The first integration focused on the raters. I wrote narrative discussions

(Creswell & Plano Clark, 2018) for each rater documenting the severity, fit and bias for their Aptis scores and on-topic decisions and the severity, consistency and construct relevance for their perceptions. The second integration focused on the test-taker responses using the same narrative discussion approach. I referred to the email exchanges to triangulate the data and confirm findings.

The convergent primary core design was the most appropriate approach for this study because the core data was collected at the same time, the quantitative and qualitative data were analysed separately and then the full data set was brought together for the final interpretation and triangulation. Figure 2 visually represents the data collection and analysis steps.

Figure 2: Research design

Part	Procedure	Product
Part 1: Data collection (QUAN + QUAL) Phase 1	30 raters listened to 45 Aptis spoken responses and provided: • 45 Aptis scores for each spoken response (QUAN) • 135 on-topic decisions (on-topic, off-topic or borderline on-topic) for the three topic relevance decisions in each response (QUAN) • 135 reasons for their on-topic decisions (how and why they made topic relevance decisions) (QUAL)	1350 Aptis ratings (QUAN) 4050 on-topic decisions (QUAN) 4050 written perceptions for on-topic decisions (QUAL)
↓		
Part 1: Data analysis (QUAN + QUAL) Phase 2	Analysis of the data (QUAN + QUAL) for each rater to develop the email exchanges	Email exchange questions
↓		
Part 1: Data collection (qual) Phase 3	Semi-structured email exchanges with raters to: • clarify decisions and perceptions • collect additional qualitative data with two follow-up questions common to all raters	Supplemental qualitative data (qual)
↓		

Part 2: **Data analysis** **(QUAN + QUAL)**	Quantitative analysis: Many-Facet Rasch Measurement to analyse Aptis scores and on-topic decisions Qualitative analysis: • Code raters' perceptions (two rounds) and writing memos (personal notes) • Second coding: 10% of data • Recode data and conduct frequency counts • Collapse (merge) codes and consolidate codes into internal criteria categories • Rater analysis	Measures of severity, fit and bias (QUAN) Codes and themes (QUAL) Five internal criteria categories that raters use to make their on-topic decisions (QUAL) Rater severity, consistency and construct relevance of on-topic perceptions (QUAL)

Part 3: **Mixed methods data analysis** **(QUAN + QUAL + qual)**	Compare quantitative and qualitative data analysis findings: • Raters: compare Aptis scores, on-topic decisions and on-topic perceptions for each rater • Test-taker responses: compare Aptis scores, on-topic decisions and on-topic perceptions for each test-taker response • Refer to email exchanges to triangulate findings	Narrative discussion on a selection of raters and test-taker responses

3.2 Participants

The participants in the three pilot studies as well as the main study were all Aptis raters and signed the consent form for participants (see Appendices 6 and 7 and section 3.3 for more information on informed consent). There were two participants in the first and second pilot studies. These were the same two participants. In the third pilot study, there were five participants. The participants in the third pilot study included the raters from the first two pilot studies and an additional three participants. There were thirty raters in the main study. The raters in the pilot studies did not participate in the main study.

The raters in the first two pilot studies were Aptis Managers, as well as raters, and paid a salary; this research was part of their full-time test validation work. The three

additional raters in the third pilot study were senior examiners working as contractors to manage rating quality assurance, as well as marking Aptis Speaking and Writing tests online at home.

For the main study, all raters were invited by email to participate (170 people) and the length of service was the only criterion used to choose participants. Raters with the longest length of service were invited to participate and all raters had at least one year of service. One year was considered the minimum length of service because within this time raters mark for a minimum of 60 hours, which gives them a good understanding of the test and scoring systems. Selecting raters by length of service for these types of research projects is normal procedure at the British Council and considered fair. Also, the evidence to include inexperienced raters, although inconclusive (Knoch et al., 2021, pp. 14-16), was not strong. For example, Lim (2011) and Sweedler-Brown (1985) both found that inexperienced raters were less accurate. Barkaoui (2011) found that experienced raters appear to be able to reflect more on the differences between the rating levels and highlight linguistic accuracy, while inexperienced raters tend to highlight linguistic errors and mark less accurately (i.e., higher levels of misfit in the data). Burton (2018) identified that raters who were experienced in rating Chinese test-takers taking an English test were more likely to consider a Chinese candidate who is off-topic to be misunderstanding the prompt and not penalise, whereas raters who were inexperienced in rating Chinese test-takers were quicker to regard off-topic responses as a memorised response and penalise. Milanovic et al. (1996) recommend using raters experienced in marking the test used in the research. I wanted to remove inexperience as a possible variable in the study, which is why I chose this approach of choosing raters with the longest

length of service. I also wanted a diverse group of raters so did not include any additional exclusion criteria.

Thirty-four raters applied for this supplementary work and 30 participants out of the 34 who asked to participate in the research were chosen based on length of service rating Aptis tests. The thirty raters in the main study were contracted raters marking tests online at home. They can work when they like and are paid according to the number of test-taker responses they mark.

Once the raters had read the participant information sheet and signed the consent form as outlined in section 3.3 (Informed consent), they completed an online background questionnaire delivered by Microsoft Forms (Appendix 9). The focus of the questionnaire was to collect information on raters' expertise in the field of language teaching and testing. The questionnaire included raters' English teaching or applied linguistics education; language teaching experience; experience rating other English language tests; and first language background of learners they have experience teaching. There were 24 female participants and 6 male participants.

I collected education in order to understand if there could be different on-topic perceptions by raters with different levels and types of education (Knoch et al., 2021, p. 20). Teaching experience was collected to explore if more or less teaching experience leads to different on-topic decisions (Knoch et al., 2021, p. 19). Experience rating other tests was collected to see if the decision-making process is affected by a preference for other types of rating scales (Fairbairn, 2015a). Language background of learners that raters have experience teaching could provide information on individual rater differences due to the rater exposure to the L1

of the test-takers (Knoch et al., 2021, p. 19). Investigation in these areas was out-of-scope for this research project but could be useful for future research.

All raters had a Certificate of English Language Teaching to Adults (CELTA), or equivalent, as a minimum professional qualification to work as an Aptis rater. Eighteen raters had a CELTA, of which seven raters had only this minimum requirement. Other education and additional education are in Table 1.

Table 1: Participant education

Professional qualification	**Number of participants**
Diploma of English Language Teaching to Adults (DELTA)	9
Graduate certificates in Teaching English to Speakers of Other Languages (TESOL) or Teaching English as a Foreign Language (TEFL)	3
Graduate diplomas in TESOL or TEFL	6
Master's degree in applied linguistics, TESOL, TEFL or language testing	10
Postgraduate Certificate in Education (PGCE) in English or languages	2
BA in languages/English/education	3
MA in English/education	3
MPhil in linguistics/ELT	2
NILE Teaching English in Pre-Primary Education	1
Doctorate	1

No raters reported 1-3 years of language teaching experience, two raters reported 4-5 years of language teaching experience, eight raters had 6-10 years' experience, and 20 raters had more than 10 years' experience.

Eleven raters reported no other rating experience other than teaching and preparing students for tests using test preparation materials. Nineteen raters had experience marking well-known international tests such as IELTS or Cambridge tests.

Most raters had taught international students from a variety of backgrounds but three raters had only taught students from one language background (Turkish, Arabic and Spanish).

Aptis raters are globally dispersed but must have a UK bank account for payment.

Aptis had been operational for 11 years at the time of data collection so the maximum experience marking this test is 11 years. There were four raters with less than 1 year experience who were not chosen for the project so the range is 1 to 11 years. Table 2 visually illustrates the rater experience on this project.

Table 2: Participant experience marking Aptis tests

Number of years marking Aptis tests	**Number of participants**
1-2 years' experience	8 participants
3 years' experience	3 participants
4 years' experience	11 participants
5 years' experience	0 participants
≥ 5 years	8 participants

The raters who were not full-time employees (33 raters) were paid a supplementary fee for this work at the Senior Examiner hourly rate for five hours of work (£17/hour). All raters in the pilot and main studies completed the work, except for one rater in the

main study who did not answer the email exchanges. I used the quantitative and qualitative data collected from this participant.

As a British Council employee, my main concern with respect to my relationship with the participants was the power imbalance between me as a researcher and the participants, particularly as I had line managed some of the participants in the past and was seen as a senior member of staff. I stressed to raters that I was genuinely interested in their opinions and that there were no wrong answers. I emphasised that the information they provided would not be sent to their current line manager or used for performance management purposes. As Kvale (2007) advises, I set the tone from the start as one-to-one professional communication with colleagues. The raters regularly participate in these types of research projects so are familiar with the process and have stated that they enjoy the experience. No problems were encountered.

3.3 Informed consent

The Examiner Network Manager sent an email (Appendix 6) to all raters (170 people) inviting them to participate in this research. Thirty raters were chosen based on the length of service rating Aptis tests (see section 3.2 Participants for a discussion on this decision).

I emailed the Participant information sheet (Appendix 7) and Consent form for participants (Appendix 8) to the 30 raters chosen for the study. The informed consent documents clarified the purpose of the study, its voluntary nature and free withdrawal, data protection and storage and the use of pseudonyms so that rater identity can be protected. Raters were issued a pseudonym by email to use when filling in the questionnaire and data collection instrument to conceal their identity.

The raters read the participant information sheet and returned the signed consent form to me by email. The pseudonyms and rater names were kept in an encrypted separate file from the data. The emails linking the raters to the data are on an encrypted server and will be deleted once this PhD is completed. I kept the emails, which include pseudonyms and rater names and email addresses, in case of lost data during the analysis. These informed consent documents were templates from University College London, which I adapted for this research project. The informed consent documents and procedures were trialled in the pilot studies and no adjustments were necessary.

3.4 Pilot studies to inform the main study

I conducted three pilot studies with the purpose of identifying 45 speaking responses for the main study and to pilot data collection instruments, interview methods and data analysis methods. I wanted to identify a suitable number of responses for the main study that would provide rich data, including cases where raters agreed on their decisions and where raters had different opinions.

For the data collection instruments, I piloted the participant background questionnaire, consent form, participant information sheet and the Excel on-topic data collection instrument used to collect rater on-topic perceptions, on-topic decisions and Aptis scores. In addition, I piloted four participant interview methods: face-to-face interview, video-conference interview, group interview, and email exchanges in order to identify the most suitable data collection method. I also used the pilot studies to train on how to use NVivo and create qualitative codes for the main study. An overview of the three pilots studies is in Table 5 in section 3.4.4 (Pilot studies summary and conclusion).

3.4.1 Pilot study 1

The purpose of Pilot 1 was to start to choose the speaking responses for the main study from the corpus of 9000 tests and to pilot face-to-face and video-conference interviews. As explained in the research context (section 1.1) and supported in the literature review, my focus was on identifying and collecting B2-level test-taker responses to the B2-level task. Lower scoring (e.g., B1 level) test-takers have more instances of topic trouble due to lower speaking proficiency (Seedhouse and Harris, 2011), which I wanted to avoid. The responses in the corpus had only been single marked so part of the pilot was also to confirm marking accuracy in the sample and eliminate any responses that had poor sound quality.

Two Aptis raters who were full-time British Council colleagues participated in Pilot study 1 as part of our regular test validation work. Both raters had been through the Aptis rater training course and had marked Aptis tests for at least one year. One rater was responsible for training new raters so was more experienced in using the marking criteria. The other rater had not done recent marking and had moved into a test development role but was responsible for the quality assurance in producing the speaking tasks. Both were studying MAs in Language Testing in the UK at the time of the data collection. They were both knowledgeable about the speaking task used in the study and available to do this work. I had a good and regular working relationship with both raters as part of the management team, so we were able to discuss marking decisions honestly. I piloted the informed consent documents using the procedures outlined in section 3.3.

3.4.1.1 Selecting responses for pilot study 1

From the corpus, I chose only the responses where a mark of 3 (B2.1) or 4 (B2.2) had been assigned by the rater on the B2 task in order to collect a sample of B2-level performance on the B2-level task. For each of the 18 countries represented in the corpus, two score 3s (B2.1) and two score 4s (B2.2) were randomly selected from one version of the test. I wanted to have a randomly dispersed and diverse sample so used responses from each country as equally as possible. All responses chosen also had a final CEFR level of B2. These responses should therefore be unambiguously at B2-level, if the rating is accurate, as the test-taker has performed at B2-level on the task and received a final CEFR level of B2 overall.

There should be 72 responses to review for pilot study 1 (18 countries x 4 responses) but for some countries the criteria could not be met (i.e., B2 score and B2 overall) and fewer responses were therefore selected. See Appendix 11 for the list of countries in the sample used in the three pilot studies.

3.4.1.2 On-topic rating perceptions

The two raters reviewed 64 Speaking Task 4 Version 1 test-taker responses using the British Council's secure file sharing software and filled in their perceptions using an Excel spreadsheet (see Table 3 for the information collected).

Table 3: Template for raters to identify on-topic responses

Rater: R001		Tick one of the three boxes: 1) if the response is fully on-topic 2) partially/fully off-topic 3) if you can't decide (borderline / ambiguous)			Tick if the sound quality is good	Notes – anything interesting about this response?
Response #	**Version**	**1) On-topic**	**2) Off-topic**	**3) Borderline/ ambiguous**	**Good sound quality**	
V1-101	1					
V1-102	1					

The raters listened to each test-taker's response. Raters are accustomed to hearing a variety of pronunciations not grouped by L1 so responses were randomised to replicate live Aptis rating. Participants in the pilot studies and main study were presented with the same randomisation order of speaking files. Raters ticked the appropriate box: a) On-topic if the response was on-topic for all three questions in the prompt, b) Off-topic if at least one of the questions in the prompt was answered off-topic or not at all, or c) Borderline if the rater was unsure if responses to any of the questions in the prompt were on-topic or off-topic. These decisions closely matched live rating by identifying on-topic and off-topic responses. Raters also ticked the box 'Good sound quality' if the response had no recording issues. Responses where raters did not tick the 'Good sound quality' box were eliminated from the sample for the main study. The data would not be useful if the raters could not hear or understand the response.

I chose written reports as a method to collect data (see 'Notes – anything interesting about this response?' in Table 3). Raters noted down how and why they reached their on-topic decisions. There were no prompts or checklists and raters completed a

report after marking each spoken performance. The purpose was to document written explanation notes to help track the topic relevance reasons.

I did not specify whether the rater should do the report introspectively or retrospectively. Introspective reports collect the information while the data collection activity takes place, while retrospective reports collect the information after the data collection activity is completed and are used when there is risk of disrupting the participant (Rose et al., 2020). Retrospective reports are not interviews but a way to collect thoughts and behaviours that are still fresh in the participant's mind following the data collection.

I instructed raters to "Write the reasons for each of the 135 decisions. Why did you decide it was on-topic? How did you make this decision?" (Appendix 10). Raters have given feedback that they take notes during live marking as a normal practice, so it was an authentic way to collect these data. However, some raters may have decided to do retrospective reports. There is a very small risk that raters who did retrospective reports could forget their decision-making process.

This method is a 'write aloud' reporting technique and was developed by Gigerenzer and Hoffrage (1995). When done introspectively, this type of reporting provides information to analyse participant decision-making processes, which does not depend on memory or how participants believe they should have thought. Participants write down their thoughts and behaviours as they make decisions so that the researcher can understand what is going through their minds (Dörnyei, 2007). This approach has been used or adapted in other L2 studies examining rater behaviour (e.g., Baker, 2012; Brown et al., 2005; Isaacs & Trofimovich, 2012). New technology is improving introspective methods via techniques such as eye-tracking

or key-stroke logging, which allows introspection without disturbing the participant (Rose et al. 2020).

Milanovic et al. (1996) recommends introspective verbal reports but it would not have been possible to complete introspective verbal reports while marking spoken responses (i.e., it is not easily achievable for most raters to speak while listening to a test-taker's response). As these were speaking tests, rather than writing tests, introspective reports would need to be written rather than verbal.

There is an assumption when using introspective and retrospective methods that raters understand and can describe their internal thought processes (Dörnyei, 2007). However, it is possible that a rater does not really know why they decided the response was on-topic or how they made these types of decisions.

Each of the Speaking Task 4 responses are two minutes of listening time by the rater. It therefore took approximately 2.5 hours per rater to review the 64 responses (2 minutes x 64 responses plus an approximately 20-minute break) and fill in the spreadsheet. There was the option to break up the work into smaller chunks. This amount of rating is in line with regular Aptis rating and conforms to industry best practice (Ling et al., 2014). No rater fatigue was anticipated or reported. The raters sent the completed spreadsheets back to me by email.

All responses without a tick for good sound quality by both raters were removed from the study. The aim was to get a sample of 15 speaking responses from the original sample of 64 responses (see section 3.4.1 for more detail on the sampling). The original sample was randomly selected and not pre-screened for sound quality, or any other criteria.

From the remaining 47 recordings that had good sound quality, responses were selected for the interview that included at least one of the following characteristics:

- Difference of opinion (on-topic, borderline/ambiguously on-topic, off-topic);
- Both raters agreed a response was borderline or ambiguous;
- An interesting comment in the notes from either rater;
- In addition, one response was chosen where both raters agreed the response was on-topic and one response where they agreed it was off-topic to collect agreement criteria.

There were 26 responses that met at least one of these criteria. Raters agreed the response was on-topic or off-topic for the additional 21 responses, which could be used as back-up.

3.4.1.3 Video conference and face-to-face interview

I next met with the two raters individually to discuss their on-topic decisions to start to understand rater perceptions of the on-topic construct and pilot two methods for data collection. As Polkinghorne (2005) notes, it is more difficult to collect evidence about human experience, which is not observable, as opposed to human behaviour. The marks which raters assign to responses and their on-topic decisions are observed behaviour but we do not know why they are making these decisions. Carefully crafted interviews might help understand this unobserved experience. A semi-structured interview schedule with a selection of responses for discussion was used to achieve this objective.

I held one interview face-to-face and the other by video conference. The interview methods were for practicality purposes but also to determine if I could conduct an interview by video conference. The main study had globally dispersed raters so it was important to understand how video conference interviewing might differ from face-to-face interviewing (pre-COVID19).

The purpose of the interview was to meet with the raters separately to probe their on-topic decisions in more depth and start to identify the 15 speech samples for the main study. It was important to set the correct tone during the interview in order for the raters to provide honest and complete feedback (Kvale & Svend, 2009). I did this by treating the interview as a regular work meeting where discussing raters and tasks is routine. The number of responses that we could review was dependent on how many samples I could get through in the allotted interview time. I reviewed 12 responses with one rater in 90 minutes and 14 responses with the other rater in two hours.

3.4.1.4 Pilot study 1 findings

Analysis of the raters' written notes and my interview notes showed that the two raters had different decision-making processes on how or why they decided a response was on-topic, different severity/leniency thresholds with respect to on-topic decisions and different preferences for analysing test-taker responses. One rater discussed test-taker responses very holistically, while the other rater was more analytical and discussed different elements of spoken speech (e.g., grammar, vocabulary, pronunciation, fluency, cohesion). One rater wrote a lot more information in the notes column, which I used to pilot NVivo' coding functionality. The other rater wrote hardly anything. My views on their decision-making behaviour emerged from

the interview comments and notes. I found this method was not very useful for collecting their perceptions, because there was too much to discuss and not enough thinking time. I abandoned this method.

The coding was done with rater 01's data using NVivo 12. The coding was not explicitly linked to the research questions and was descriptive but coding categories started to emerge, for example:

- Internal criteria: The rater perceived when a test-taker answered the three questions in order (a linear response) and when they integrated their responses into a long turn (see Appendix 2 for the task prompt for version 1). They also mentioned poor test-taker technique.

- Task prompt: In this task there is a picture to help put the task into context. The picture in version 1 shows a boy on a bicycle and the task is about outdoor activities. Most test-takers discussed cycling even if they had never done it before or disliked the activity. The rater noted this task effect. Another task effect was that one question could be answered with a yes/no response, rather than with a more elaborated response, and still be on-topic.

- Rating criteria: Some test-takers were clearly at a high level but spent the full time answering only one question. The rater noted these unbalanced responses. Other lower-level test-takers had better strategies and seemed to have prepared for the test, and in one case may possibly have had a pre-prepared answer. The rater noted how the language level and the on-topic rating requirements in the rating scale did not always align in the scoring.

The face-to-face and video conference approaches worked equally well but I was not entirely satisfied with the methods. Our team works globally and remotely, which explains why doing the interview online was not problematic. I would have been confident to do the interviews by video conference in the main study. However, there was not enough time to analyse the test-taker responses and rater perceptions in any depth to really identify issues. The interviews felt superficial and only really clarified opinions. The face-to-face rater was very quiet during the interview and had written fewer notes and so I was speaking more than the participant. The online rater dominated the interview with lots of notes and lots to say and I spoke very little. It was a good way to gain a basic understanding of their overall view of the on-topic construct but not to explore these views in any depth. It felt like an unbalanced data collection strategy. A grounded approach and multiple follow-up interviews could have enhanced this data collection method. I needed time to prepare follow-up questions once I started to better understand their perceptions. Follow-up interviews at each stage of the analyses might have enhanced the findings. Multiple interviews, however, were not financially feasible for 30 raters.

I therefore conducted a second pilot with a new version of the task. There had been discussion about the content of the task (riding a bicycle) and I wanted to determine if the theme of the task had any effect on topic relevance decision-making by raters. I also decided this time to conduct a group interview and bring the two raters together in order to stimulate a deeper discussion.

3.4.2 Pilot study 2

The goal of Pilot 2 was to examine how a group interview would work as a method for collecting data. I chose a group interview to try to stimulate a deeper discussion

with the two participants, which was missing in the individual interviews. I thought it might lead to richer data and confirm or identify a particular tone or strength of feeling about topic relevance. I also wanted to look at possible task effects using a second version of the task.

3.4.2.1 Selecting responses for pilot study 2

I used the same procedures as Pilot 1 to collect the sample from the new task (version 2 – see Appendix 2 for the task) and the same data collection Excel spreadsheet (Table 3), as well as the informed consent procedures (section 3.3). From the possible sample of 72 responses, 54 responses met the B2 criteria. Of the 54 responses, 37 remained after I removed the poor sound quality responses.

3.4.2.2 Group interview

Instead of individual meetings, I held an online group interview with both raters to examine if and how "collective sense-making" (McKinley & Rose, 2020, p. 293) could take place and to stimulate a deeper discussion with the two raters. As there were only two participants, I termed the interaction a 'group interview' rather than a 'focus group'. Focus groups should include 6-10 participants and are more formal, directive and structured than group interviews (Morgan, 1996). I thought a group interview might lead to richer data and confirm or identify a particular tone or strength of feeling about topic relevance. The raters had very different decision-making processes, feedback and approaches to giving feedback so there was some risk in holding a group interview. Subjective bias can occur if a rater or the researcher dominates, silences or interrupts the discussion and causes a rater to change their mind about their decisions (Johnson & Onwuegbuzie, 2004). According to McKinley

and Rose (2020), the group interview becomes an additional variable in the study and requires an analysis of the group discussion.

3.4.2.3 Pilot study 2 findings

My initial concerns about holding a group interview were confirmed in pilot study 2. One rater dominated the discussion and might have been influencing the other participant. The raters were gaining consensus, which meant that they were changing their minds. I realised that I was more interested in finding out what participants think on an individual level and not in building consensus. Moreover, the process was very similar to the individual interviews and the data collected was again superficial, except now there was a group interview bias variable. Group interviews were unsuccessful as a method to confirm and draw conclusions and I abandoned this method.

In pilot study 2, one rater added extra columns in the data collection spreadsheet (Table 3) for additional information. They added the perceived gender of the test-taker, the Aptis score they would assign, as well as key vocabulary they used as evidence of on-topic responses. This information provided indications of what raters might perceive to be important when making the on-topic decision for me to consider in a future iteration of the data collection instrument and data analysis. Gender of test-takers could be a source of construct irrelevant variance but a gender study was outside the scope of this project. The constructs for topic relevance decisions and vocabulary are not independent; it would be difficult to talk about a topic without the required vocabulary. This finding corroborates research into the closeness of topic relevance and vocabulary constructs (for example Lumley, 2005; Milanovic et al., 1996; North, 2003; Sasayama & Norris, 2022). However, it should be noted that a

test-taker can still achieve a task's goals efficiently and speak on-topic without necessarily developing the topic or knowing anything about the topic, as noted by Pallotti (2009). This data also gave me an insight into how raters might have difficulty separating the on-topic decisions from the Aptis score.

Pilot study 2 made me aware that there were task effects as a result of different versions. The task 2 prompt was about a long journey and both raters were quite particular on what constituted a 'long journey'. One rater considered a response off-topic if the journey was not long enough. The other rater focused on the 'journey' and marked off-topic if the response did not talk exclusively about the 'journey' and instead talked about the destination or holiday. This rater raised a good point that 'journey' is more abstract and more difficult to describe than a holiday, which uses lower-level vocabulary. Therefore, there were test-taker responses where the two raters in this pilot study made the same off-topic decision but for different reasons. This insight provided evidence that raters could have different perceptions on what constitutes speaking on-topic, which, in the context of the Aptis speaking task 4, could influence the on-topic decisions and Aptis task scoring. It is possible that a different rater would not perceive either of the points raised by these two raters and decide the response is on-topic and award a higher Aptis score. Therefore, this is a scoring validity question. This finding is consistent with the literature on rater variability raised in section 2.3 in the literature review.

In pilot study 1, the raters had noted the task effect of the bicycle photo impacting on test-taker responses, while in pilot study 2 the raters had disagreed on their definition of a 'long journey'. The task effects found in pilot studies 1 and 2 and the limitations mentioned above in using the group interview method made me decide to pilot a

third version using email exchanges for the data collection method. I believed that task effects were an important part of the on-topic decision and I wanted to have as many question prompts in the main study as possible.

3.4.3 Pilot study 3

The aim of the third pilot study was to pilot email exchanges and a third version of the task. I also added three raters to the participant sample for a total of five raters, including the two raters from pilot studies 1 and 2. I decided to pilot email exchanges as a way to improve this part of the qualitative data collection. I thought email exchanges would give raters and me more time to think before responding and therefore provide richer data. We communicate almost exclusively by email about rating issues so it would also be a more authentic data collection method.

3.4.3.1 Participants

I added raters to the sample who were similar to the cohort in the main study to understand any issues that did not come up in the first two pilot studies. The first two raters were close colleagues who worked in management roles. We were accustomed to reviewing raters and tasks together, so the work was familiar and comfortable. The three raters I added to the third pilot study were experienced Aptis senior raters working as contractors paid an hourly fee to support and mentor junior raters. The five raters piloted version 3 (see Appendix 2 for the task version). The senior raters also piloted the invitation to participate email, participant consent form, information sheet and participant background questionnaire (see Sections 3.3 and Appendices 5-9 for more information on these recruitment and informed consent documents and procedures). The main difference in the administration of these documents, compared to the main data collection process, was that the invitation to

participate was sent only to the 15 senior raters and not the full examiner pool. Also, I used SurveyMonkey instead of Microsoft Forms to collect the participant background questionnaire. The new raters were not British Council employees and we were not allowed to use Forms with external contractors at the time. The three new raters reviewed versions 1, 2 and 3 and filled in the Excel spreadsheet. This additional data provided more information for selecting the sample for the main study.

3.4.3.2 On-topic perceptions research instrument

For pilot study 3, I made changes to the on-topic perceptions research instrument by adding to the notes column more explicit information asking how and why they made their on-topic decision (see Table 4). In the first two pilot studies, the raters had approached the notes section in very different ways, so I added these clarifications as a way to aid raters to pinpoint the evidence for their decisions and to help standardise the information. Senior raters filled in their decisions for each response at a time of their choosing and at their own pace. The senior rater names and test-taker responses were again pseudonymised (e.g., R003 and Response V1-101).

Table 4: Updated template for pilot study 3 data collection

Rater: R003		**Put an X in one of the three orange boxes:** **1) if the response is fully on-topic** **2) partially/fully off-topic** **3) if you can't decide (borderline / ambiguous)**			**Put an X if the sound quality is good**	**Why did you decide it was on-topic?** **How did you make this decision?**
Response #	**Version**	**1) On-topic**	**2) Off-topic**	**3) Borderline/ ambiguous**	**Good sound quality**	
V1-101	1					
V1-102	1					

3.4.3.3 Semi-structured email exchange

I contacted the raters by email to clarify their decisions and explanations, to probe into their on-topic decision-making perceptions in more depth and to collect two follow-up qualitative data. In addition to the evidence from the previous two pilot studies, research on email exchanges contributed to choosing this data collection method. According to Ratislavová and Ratislav (2014), participants are generally more focused during email exchanges and the data are denser, more structured and more explicit compared to face-to-face interviews, video interviews or group interviews. Raters have more time to think, reflect and consider answers, which helps them engage in more careful communication. Baker (2012) highlights the benefits of write-aloud protocols, as opposed to speak-aloud protocols, in that it does not disrupt normal marking, data does not need to be transcribed and interpreted and the method may require less training than verbal protocols. There is a risk that what they write might not be what they are actually thinking and, with more time to think, they might give an enhanced account to please the researcher. I also had more time to consider my questions and replies in the subsequent follow-up email exchanges so I was more satisfied with the probing questions.

One issue with using email exchanges is that participants might be less able to explain themselves in writing than in speech. However, these raters are accustomed to discussing marking issues with management staff by email so the format was authentic and familiar for raters. Also, they are expert users of English so could effectively communicate. The key to obtaining rich data according to Meho (2006) is to make sure the emails lack ambiguity, provide space for additional follow-up questions and encourage a relaxed communication style. Hawkins (2018) notes that

email exchanges can also lack some of the social cues that contribute to an understanding of the rater's experience.

To create the email exchange, I reviewed each rater's data in the *On-topic perceptions and scoring data collection instrument* (Table 4) individually as recommended by Patton (1980) and Teddlie and Tashakkori (2009). Patton (1980) discusses case study approaches while Teddlie and Tashakkori (2009) describes this process as analytic induction and involves looking for patterns in a rater's decisions. I analysed each rater's data to understand patterns of decision making and used this analysis to develop questions for the follow-up email exchanges. The analysis was a high-level overview of patterns and abnormal decisions in the quantitative data (e.g., marking a 4 / B2.2 but with off-topic decisions) and patterns in the qualitative data (e.g., interesting words and approaches to decision-making). These were mini evaluations with the purpose of understanding the rater holistically in order to create the email exchange questions for further data collection.

I also did not consider a stimulated recall verbal protocol for my research for similar reasons for the verbal protocol method mentioned earlier.

3.4.3.4 Pilot study 3 findings

The first two raters from the first two pilot studies had been aware of the study for some time and understood the objectives. It became clear during pilot study 3 that the three new raters had different understandings of expectations. They also found it difficult to only talk about the on-topic part of the construct.

- One rater thought they wouldn't get to relisten to test-taker responses during the email exchanges so took lots of notes.

- One rater transcribed all the speaking files and put those in the notes section and then highlighted the relevant parts that showed the evidence for an on-topic response.

- One rater was worried about their decisions and wanted to know if I thought they were harsh or lenient in their on-topic decision-making.

Discussing the responses with the raters was easier with email exchanges because I did not need to think of the next question while actively listening during a face-to-face interview, video conference interview or group interview discussion. The data were also richer and more considered, did not go off-track and the discussion was more focused on the research questions. Raters are globally dispersed (raters mark online and are not located in a hub) so this method was practical.

I termed these as ‘email interviews’ for the pilot study but it became clear that this type of data collection is not really an interview. The interactive nature of synchronous communication (written or spoken) is missing and many of the elements that occur when speaking with someone are missing. There can be miscommunications that cannot be quickly clarified. The communication is more formal and well thought through. They are convenient and practical but lack the shared focus of an interview. There could be distractions and different amounts of time spent on answering emails and the researcher and participants are having different experiences at the time of writing. I decided to term this method as asynchronous email exchanges.

For this pilot study, I linked the NVivo coding to the research questions. I started with codes for the research questions (rater internal criteria, rating scales and task

prompt) and then coded anything that referred to the research questions and developed sub-categories under the research questions. At this stage, I was still learning how to use NVivo and experimenting with different kinds of coding structures.

The NVivo coding highlighted an issue. There were responses where the test-taker did not answer one of the questions. The response was not off-topic but simply missing. I decided to remove these responses from the sample for the main trial if all five raters agreed the response was missing because this data would be less informative for my research questions. I wanted to understand the reasons for rater on-topic decisions and, without anything to evaluate, the data would not be useful.

In the pilot studies, I asked raters to decide if the full response was on-topic and it became clear that some raters were making decisions based on how many of the three prompts were answered on-topic. This finding is unsurprising as this is the construct in the rating scale. However, raters commented on different aspects of the response and not always on each prompt.

I decided that for the main data collection that I would ask raters to provide separate comments on all three prompts. For some raters it was an unnatural way to think about marking, but I felt that I needed to probe this decision more explicitly to understand their values about the on-topic construct in spoken performance. I was interested in how they explain their on-topic decisions for each of the three responses. Although this method of data collection was more experimental, I believed it would better help address my research questions. The research questions were expanded from analysing ‘task effects’ to looking into the impact of the three ‘prompts’ and three ‘versions’.

Another benefit of separating the three on-topic decisions in the data collection was that it allowed me to collect quantitative data to understand rater severity and reliability across prompts and versions. Each response has three prompts so there were 135 on-topic decisions to make (15 responses x 3 versions x 3 prompts). With 30 raters and 135 decisions, this sample size was considerably expanded and sufficient to allow the decisions to be analysed quantitatively using Many-Facet Rasch Measurement (Eckes, 2015; Green, 2013).

3.4.4 Pilot studies summary and conclusion

See Table 5 for an overview of the three pilot studies including the data collection and data analysis methods, key findings and implications for the main study.

I have followed a similar methodology to Brown et al. (2005) by collecting rater decisions without any criteria so that I could investigate their unguided perceptions. Key differences are that I used three tasks instead of five and all my tasks were the same task type. I also only focused on one CEFR level (B2) whereas Brown et al. (2005) looked at five levels. There was a larger test-taker response sample in the Brown et al. (2005) study but fewer raters. I also used written, rather than verbal, reports and a fully crossed design.

Table 5: Overview of pilot studies

	Pilot Study 1	**Pilot Study 2**	**Pilot Study 3**
Data collection	Identify 15 B2-level responses for main study from version 1 Two participants Identify if responses are on-topic, partially / fully off-topic or borderline / ambiguously on-topic	Identify 15 B2-level responses for main study from version 2 Same two participants Same Excel data collection instrument to collect on-topic decisions	Identify 15 B2-level responses for main study from version 3 Five participants, including the first two participants Updated Excel data collection - included

	using an Excel data collection instrument Notes of anything interesting about the responses Identify poor sound quality responses Video conference and face-to-face interviews	Notes of anything interesting about the responses Identify poor sound quality responses Group interview	questions of 'how' and 'why' they made their on-topic decisions Semi-structured asynchronous email exchanges
Data analysis	Analytic induction / mini evaluations of Excel data Review of rater 01 notes during an NVivo training course and created initial codes	Analytic induction / mini evaluations of Excel data Continued to practice coding with NVivo	Analytic induction / mini evaluations of Excel data NVivo coding linked to the research questions
Findings	Different decision-making processes between the two raters No issue with online vs face-to-face interviews Not enough time to think and discuss perceptions – interviews felt superficial Unbalanced rater data due to different decision-making processes Photo effects Task effects	Group interview dominated by one participant Same superficial finding from pilot study 1 Group interview bias variable noted Consensus not desirable Raters might have difficulty separating the on-topic decision from the overall Aptis score Vocabulary and on-topic decisions finding Task effects	Difficulty for new raters to speak only about the on-topic part of the construct Email exchanges worked well – rich and considered data Removed test-taker responses where one of the responses was not answered (rather than off-topic)
Implications for main study	Online and face-to-face interviews abandoned	Group interview method abandoned	Email exchanges yielded authentic and rich data and was chosen for the main study Decided to split the decisions from one decision on the full response to three decisions on each of the three question prompts Finalised 45 speaking responses

3.5 Data collection procedures for the main study

Following the three pilot studies, the instruments for the main study were finalised, email exchanges were chosen as the best method for follow-up supplementary data collection, 45 speaking responses were chosen for the main study (15 from three versions of the task) and the research questions were confirmed. The informed consent procedures and documents were also finalised.

I followed the informed consent procedures in section 3.3, which I had trialled in the pilot studies. Once the Participant information sheet (Appendix 7), Consent form for participants (Appendix 8) and the Participants background questionnaire were completed (section 3.2 and Appendix 9), I sent the data collection instrument, test-taker responses and the email instructions (Appendix 10) to the raters. The next sections provide the data collection procedures for the main study.

3.5.1 Quantitative data collection

The first quantitative data collected were the on-topic decisions (columns B, D and F in Table 6). There are three questions in the task prompt and raters were asked to choose from a drop-down list whether each response was on-topic, off-topic or borderline/ambiguous.

Table 6: On-topic perceptions and Aptis scoring data collection instrument

	A	B	C	D	E	F	G	H
1	Response #	Prompt 1	Why did you decide it was on/off topic? How did you make this decision?	Prompt 2	Why did you decide it was on/off topic? How did you make this decision?	Prompt 3	Why did you decide it was on/off topic? How did you make this decision?	Aptis score
2	V1-101							
3		On-topic						
4		Off-topic						
5		Borderline/ambiguous						
6								

This method provided 135 on-topic decisions (45 responses x 3 prompts). These on-topic decisions were converted to a rating scale for analysis:

On-topic = 3

Ambiguous/borderline = 2

Off-topic = 1

Three versions of the task were used in the main study (15 responses per version).

The second quantitative data collected were the 45 Aptis scores in column H, one for each test-taker response on a scale of 0-6.

Linacre (2002) and Eckes (2015, p. 132) recommend the sample size should have at least 30 observations per element and 10 observations in each rating scale category. The elements in this study, the test-taker responses (45) and the raters (30), meet the minimum of 30 observations per element. There were 1350 Aptis scores (45 responses x 30 raters). With three prompts per response there were 4050 on-topic decision observations (45 responses x 30 raters x 3 prompts). Also, according to Linacre (2002) and Eckes (2015, p. 132), the sample size should have at least 10 observations for each rating scale category. In this research, there are two categories of expected Aptis scores (B2.1 and B2.2) requiring a minimum of 20 observations and three categories in the on-topic decisions (on-topic, borderline and off-topic) requiring a minimum of 30 observations. This minimum observation count was achieved and is further discussed in the quantitative analysis results in section 4.1.

3.5.2 Qualitative data collection

The main qualitative data are in columns C, E and G in Table 6. I asked raters for an explanation for their 135 on-topic decisions with two questions to guide their written response: *Why did you decide it was on-topic? How did you make this decision?* The purpose was to understand how raters make their on-topic decisions and also to develop the follow-up semi-structured email exchange questions. The data collected is separate for the three prompts, which was a change from the pilot studies. Raters were allowed to listen to the responses as many times as they wanted and there were no time constraints; I wanted raters to take as much time as needed to make their decisions and write their perceptions. Raters completed and returned the spreadsheet with their perceptions and decisions to me by email following the same procedures in the pilot studies.

The supplementary qualitative data collection was the follow-up semi-structured email exchanges with each of the 30 raters to clarify perceptions and collect additional qualitative data. These email exchanges followed the procedures in pilot study 3 (see section 3.4.3.3 for a literature review on email exchanges) and took place as soon as possible after receiving the *On-topic perceptions and scoring data collection research instrument* (Table 6) from the raters. In order to quickly exchange emails with raters, the 30 raters were drip-fed into the research in monthly batches of 5-7 raters so that the workload was manageable and the amount of time between receiving the data from the raters and sending the email exchange questions was very similar (about one week). See Appendices 12-14 for the email exchange templates and an example of the data.

The email exchanges were semi-structured in that some parts of the email had questions that were common to all raters, while some parts of the email were unique to each rater. The unique questions for each rater focused on the patterns and outliers (e.g., unexpected Aptis scores, on-topic decisions or on-topic perceptions) found in each rater's Excel spreadsheet (e.g., Why did you score the response a 4 when you also decided two prompts were off-topic?).

There were two purposes for the email exchanges. The first purpose was to triangulate, clarify and fine-tune the mixed methods interpretation. I compared the quantitative and qualitative analyses for each rater and then referred to the email exchanges to identify if there were any further explanations for the findings. I documented any changes in my views on their severity, reliability and construct relevance. I was cautious about making changes because this part of the data collection was not standardised. Some raters reflected and were more engaged than other raters. One rater did not respond to the email exchanges.

The second purpose was to understand if raters' reported rating strategy or perceived emotion when making topic relevance decisions which might affect topic relevance perceptions and decisions or Aptis scoring. I collected this information in the follow-up email exchanges so that there was no confusion with the core data collection activity (i.e., Aptis scores, on-topic decisions and on-topic perceptions). I asked two questions to all raters:

- What is your marking strategy? How do you normally go about marking Speaking Task 4?

- Was there anything you felt really strongly about? One of the down-sides of email interviews is that it can be difficult to understand emotion so if there is anything where you felt any type of emotion when marking the on-topic criteria [on-topic rating requirements], please let me know. It might help me prioritise decisions as important or not so important.

The first question was to find out how raters normally mark this type of speaking task to further understand how the on-topic decisions fit into their reported normal marking behaviour. The procedure to collect this data was not as comprehensive as Papajohn (2002), who collected visual concept maps, but could help explain findings in the mixed methods interpretation.

I asked the second question about their strength of feeling to attempt to overcome one limitation when using email exchanges. Visual and non-verbal cues and strength of feeling are missing.

One limitation of this method is that I had to choose a reasonable number of questions to ask. Raters were paid for one hour to answer email exchanges. I was also asking the questions as I received the Excel spreadsheets and viewing raters independently; I had not analysed the full data set.

3.6 Data analysis procedures for the main study

The data analysis used a fixed mixed methods convergent primary core design as explained by Creswell and Plano Clark (2018). The quantitative and qualitative data analysis for the main study took place separately and independently. For the quantitative data, I conducted Many-Facet Rasch Measurement of the Aptis scores and on-topic decisions using FACETS software. The purpose was to understand

rater severity, reliability and bias for the Aptis scores and the on-topic decisions. For the qualitative data, I performed comparative thematic analysis of the on-topic written perceptions using NVivo and Excel software. The mixed methods interpretation integrated and triangulated the objective on-topic decisions and Aptis scoring, the subjective reasons for topic relevance decisions and the email exchanges for a complete interpretation of the data and to answer *RQ2: What is the relationship between the internal criteria raters reportedly use to make their on-topic decisions and raters' Aptis scoring and on-topic decisions with respect to rating severity, reliability, consistency and construct relevance?* The next sections contain the detail of the data analysis procedures for the quantitative, qualitative and mixed methods analyses.

3.6.1 Quantitative analysis procedures

The first step was to analyse the Aptis scores and the on-topic decisions, which were analysed in a fully crossed design using Many-Facet Rasch Measurement (Aryadoust et al., 2020; Eckes, 2015; Green, 2013; Linacre, 1989, 2018a). Many-Facet Rasch measurement (MFRM) analysis is used for data that derives from human decision-making using a rating scale. There must be at least three categories in the rating scale for the data to be considered polytomous (as opposed to dichotomous which only has two categories, e.g., yes/no) and suitable for MFRM (Penfield, 2014). MFRM examines the probability of each score point in a dataset with respect to the data as a whole and evaluates the extent to which the raters' decisions (on-topic decisions and Aptis scores in this research) are consistent with the decisions of the other raters in the sample. For example, was an on-topic decision or Aptis score expected? Rater decisions can then be measured and

compared on a ruler or scale using logits (Bachman, 2004; Eckes, 2015; Green, 2013; Knoch et al., 2021; McNamara et al., 2019). Logits are the entries on a ruler or scale that have been transformed from the data, taking into consideration the probability of obtaining a particular rating given parameters for all variables in the study, for example task type (prompt and version), rater or rating scale (Myford & Wolfe, 2003).

I considered generalisability theory for my quantitative research into how and why raters decide if a spoken response is on-topic but decided to use Many-Facet Rasch Measurement. I was not primarily looking for measurement error and score reliability on a large random sample of test-taker responses. Instead, I wanted to specifically identify rater severity, reliability and bias with respect to on-topic decisions and Aptis scoring that I could compare with the on-topic perceptions in the mixed methods interpretation and triangulation. Many-Facet Rasch Measurement was the best software to give me this information.

I used the computer program FACETS (Linacre, 2018a) to conduct the MFRM analysis. I chose this software because it provides not only rater severity and leniency (which could be determined using Excel) but also fit, unexpected responses and bias statistics and other detailed individual and group statistics about the raters and the test-taker responses. The software plots the raters onto logit scales according to their rating severity. More severe raters mark lower than the group while lenient raters mark higher than the group. The software can also identify inconsistent raters who sometimes have severe and sometimes lenient Aptis scores and on-topic decisions, or raters who have bias towards some aspect of rating. An example of bias could be a rater who consistently marks a type of test-taker or type of test

question in a way that is different to the other raters in the sample. I analysed the on-topic decisions overall, for only the three different prompts and additionally for only the three different versions of the task using this software. I also analysed the Aptis scores.

Scoring decisions that are unexpected can also be flagged in the 'fit' statistics for further analysis. Fit statistics provide information about the variability in the ratings of individual raters (Eckes, 2015). These statistics compare the observed scoring to the predicted scoring in the measurement model. The measurement model uses estimated rater severity, ratings of other raters, estimated difficulty of the question prompt with respect to rating criteria and estimated performance level of the test-taker (Myford & Wolfe, 2003). The observed ratings are compared to the expected ratings and calculated as residuals. The absolute residuals are standardised, squared and averaged across raters to provide summary fit statistics.

A normal summary fit statistic is 1.0. Raters with inconsistent decisions have fit values above 1.0, which means they have variation in their rating that is inconsistent with the sample of raters (large absolute residuals). These raters may sometimes be severe and sometimes lenient compared to other raters in the sample and are considered to be 'misfitting'.

Raters with infit (or outfit) values less than 1.0 have less variation than expected and are 'overfitting'. Their ratings are too predictable or provide redundant information. These raters might be exhibiting a central tendency or halo effect or be colluding with other raters by providing the same scores (Eckes, 2015, p. 77). Overfit is out of scope for this research.

Fit statistics can also be analysed for the test-taker responses. Misfit statistics identify which test-taker responses were scored inconsistently by raters. Test-taker responses with a range of on-topic decisions and/or Aptis scores can be compared with the topic relevance perceptions in an attempt to understand the reasons for the decision-making differences.

The fit statistics report infit and outfit. Outfit is the average of the squared standardised residuals and sensitive to individual unexpected ratings, i.e., outfit includes outliers, while infit weights the outliers less heavily (Myford & Wolfe, 2003). In my research, outliers are important so I have included both in my analyses.

McNamara (1996) proposed that raters with fit statistics between 0.7 and 1.3 are rating with appropriate variation while other researchers have set the thresholds at 0.5 and 1.5 (Eckes, 2015; Green, 2013) or 0.6 and 1.4 (Wright & Linacre, 1994). The appropriate threshold is dependent on the purpose of the analysis and often linked to the stakes of the test (Myford & Wolfe, 2003; Wright & Linacre, 1994). If life-changing decisions are to be made about a test-taker, rater reliability is more important and misfit statistics thresholds should be stricter. Wright and Linacre (1994) also note that these thresholds can change for 'off-target' analysis, such as in my research, which can produce idiosyncratic fit statistics due to a non-random sample.

The choice of fit statistic threshold is also dependent on sample size. The larger the sample size of raters or test-taker responses, the smaller the variance of the fit values. The upper limit for the infit statistic (without outliers) should be equal to $1 + 2/\sqrt{N}$, while the outfit (with outliers) should be $1 + 6/\sqrt{N}$, where N is the number of raters or test-taker responses (Eckes, 2015, p. 78; Wolfe & McVay, 2012). For the 30 raters in my research, the upper limit should be 1.365 for infit and 2.095 for outfit. For

the 45 test-taker responses, the upper limit should be 1.298 for infit and 1.894 for outfit. I chose a cut-off of 1.3 for both infit and outfit in line with McNamara (1996); however, as noted, this is a non-random sample so fit statistics thresholds should be used with some caution. I also reviewed the standardised fit statistics, which determines the significance of the infit or outfit findings. A value outside the range of +/-2.0 means that the finding is significant.

There is a small number of missing quantitative data (Aptis scores). Baker (2012) encouraged raters to mark two scores if they were unsure of their decision as a way to identify avoidant and dependent raters. The hypothesis was that the decision to give two scores is not making a decision (avoidant) and would need to be adjudicated by senior raters (dependent). I triangulate missing data in the mixed methods interpretation as an additional check to determine if there is any evidence of avoidant behaviour.

3.6.2 Qualitative analysis procedures

The qualitative analysis examined the on-topic written perceptions from raters using inductive comparative thematic analysis at the semantic level using NVivo and Excel software. The qualitative analysis took part in five steps. First, I analysed and compared the rater perceptions for each test-taker response to identify patterns and themes in the data. I coded the data in NVivo with themes that emerged from the words used by the raters. This first analysis across the test-taker responses was in two rounds, with the second round fine-tuning the emerging themes. I also collected memos, or personal reflections, as I coded. Second, 10% of the data was second coded using the codes from the second round of coding in order to increase confidence in the codes. Third, I recoded the rater perceptions using findings from

the second coding and conducted frequency counts to document the importance of codes. Fourth, I grouped the codes and identified five internal criteria categories that raters use to make their on-topic decisions. Fifth, I analysed and categorised raters' by the severity, consistency and construct relevance of their on-topic decisions.

3.6.2.1 Coding rater perceptions of test-taker responses

The first type of analysis of rater perceptions was inductive open coding. I started by using elements of grounded theory (Glaser, 1978, 1992; Glaser & Strauss, 1967) and coded the rater perceptions data line-by-line. I did not create the themes a priori or use a pre-existing coding framework. The codes emerged from the data and were directly linked to the research questions (Boyatzis, 1998; Patton, 1980; Saldana, 2016).

The coding was at the semantic level (Braun & Clarke, 2006) in that I was not looking for anything beyond the literal meaning of what the raters had written. I did not attempt to make any assumptions about the data or do any discourse analysis of what the raters had written. We cannot know what raters think, mean, feel or believe with their perceptions. We can only see what they wrote and make assumptions on what they have articulated. The codes were descriptions of raters' perceptions (Saldana, 2016) with the purpose of identifying what the text meant. The closer the codes are to the raw information increases the likelihood of inter-coder reliability (Boyatzis, 1998), which is why I chose this approach.

3.6.2.1.1 **Coding for themes (round 1)**

I coded each written perception (4050 data items = 30 raters x 45 responses x 3 prompts) in its entirety to keep the perception in context and allocated perceptions to multiple codes. I did not segment the data in any way. Perceptions ranged from one

word to several sentences. Because I coded the entire perception, the text could be allocated to several codes. For example, a rater may use a quote as evidence of the response being on-topic (quote code). They may also mention that it was too brief (amount of response code). Saldana (2016) calls this ‘simultaneous coding’ and cautions that applying more than one code to a chunk of text could be seen as indecisiveness. This type of coding may suggest no clear or focused research purpose. I justified the allocation of perceptions to multiple codes because I wanted to understand all the possible reasons raters might use for making their on-topic decisions in this first round of coding. I did not want to fragment the data too early in the process.

I deleted, expanded, refined, defined and consolidated the codes in an iterative manner into categories linking together similar and connected codes with inclusion rules for each code (Bazeley, 2013; Boyatzis, 1998; Saldana, 2016). My data collection method changed from the pilot studies in that I was now collecting perceptions for all three prompts separately, rather than for the full response. Some of the codes from the pilot studies were no longer relevant and new codes emerged.

In this first round of coding, I coded in order by response, prompt and version. I first coded each rater’s perceptions for prompt 1 for version 1, then prompt 1 in version 2 and then prompt 1 in version 3. Then I moved on to prompt 2 and then prompt 3.

I created separate data sets for the three prompts as the codes were different depending on the prompts.

Prompt 1: Tell me about… (an activity or event in the past)

Prompt 2: How did you feel about… (the activity or event in the past)

Prompt 3: Abstract question on society related to the activity or event

I wrote memos for prompts and versions on any patterns or difference of perceptions that were noticeable as I coded and took quotes to support my perceptions. There were 135 analyses (45 test-taker responses x 3 prompts).

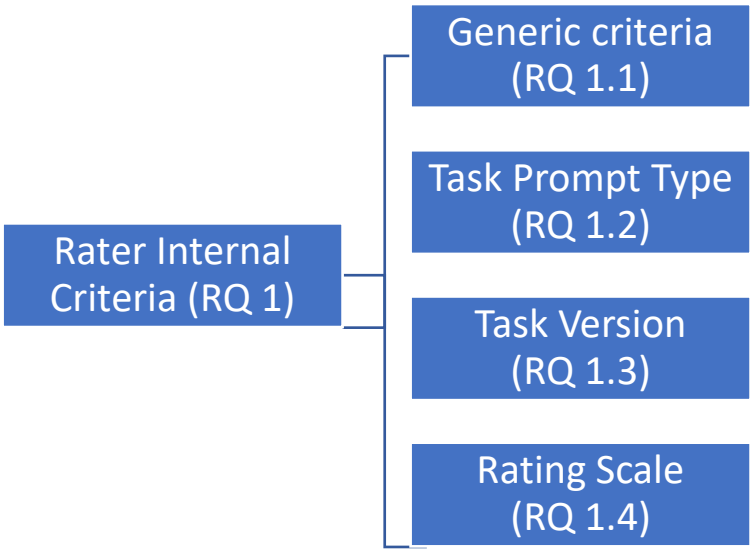
The first round of coding gave me a good understanding of the data and emerging themes and allowed me to develop codes.

3.6.2.1.2 **Coding for themes (round 2)**

My research questions were fine-tuned after the first round of coding and I did a second round with the updated research questions. As the analysis progressed in round one of coding, it became clear that all perceptions were essentially rater internal criteria. Rater internal criteria became the main research question with the prompt and rating scale becoming sub-research questions. I also found as I worked through the data that there were different perceptions depending on the task version, so I added a research question about versions (*How do the task versions from different forms of the test affect raters' decision-making?*). I also noticed that some codes were used across all prompts and versions so I added a generic category (*What generic criteria do raters report using that are common across all speaking test prompts and versions used in this research?*). I fine-tuned the research questions and recoded the data.

See Figure 3 for the main categories in the coding structure linked to the research questions in round 2 and see section 2.4 for the research questions. The codes are in Appendix 15 and detailed below.

Figure 3: Categories in second round of coding



The 'generic criteria' codes included rater internal criteria codes of perceptions that were used across all prompts and versions. These included the following areas:

- Amount of test-taker response (e.g., brief, not enough detail, depth, detailed, at length)
- Clarity of test-taker response (e.g., clear, clearly, unclear, implied, explicit and fluency)
- Picture mentioned
- Words in prompt used as evidence
- Examples from test-taker response used for evidence
- Quotes from test-taker response used for evidence
- Format: Layout of the test-taker response mentioned (e.g., integrated response, well structured)

- Pre-prepared test-taker response
- No evidence of how or why the rater made their decision (e.g., “The speaker gives a relevant response” R14)
- Test-taker response is off topic (used to collect perceptions for off-topic decisions)

The ‘prompt’ category included codes that were specific only to each prompt type. Included in these codes were rater perceptions that referenced the prompt type, examples related to the prompt type and issues around the prompt type.

- For prompt 1, the coded perceptions by some raters focused on test-takers discussing multiple activities, journeys or buildings, rather than one activity as the prompt requires (e.g., “Doesn’t really go into one activity, mentions several” R06).
- For prompt 2, the coded perceptions were used when the rater does not correctly identify a feeling (e.g., “Doesn’t talk about feelings but describes how it was hard and overcame the obstacle and learnt from it” R35).
- For prompt 3, coded perceptions included where a test-taker does not generalise and, for example, continues to talk about a personal experience (e.g., “Mainly speaks about personal experience only; at very end starts to speak more generally but doesn't answer prompt” R31).

The ‘task versions’ category included codes that collected perceptions that mentioned words in the prompt specific to the task version or theme of the questions.

- For version 1 (outdoor activities as a child), the coded perceptions were how raters perceived different types of outdoor activities.

- For version 2 (long journeys), the coded perceptions included raters perceptions of if trip, holiday or destination could be considered a journey. Perceptions also included what constitutes a suitable length, time or distance to be considered a 'long' journey.

- For version 3, (tall buildings), there were no codes that emerged at this point in the study. I coded perceptions where raters noted that test-takers did not refer to the name of the tall building. However, I coded these as prompt type, not version.

The 'rating scale' category included codes that collected perceptions that mentioned language, for example:

- Verb tenses (e.g., talking in the wrong tense)

- Vocabulary or lexis

- Other language (e.g., grammar - not verb tenses; personal or impersonal errors; fluency; discourse markers; coherence; and cohesion) or where raters mention "overall language" without specifics

I collected memos as I coded the rater perceptions. These were bullet point notes on patterns, differences of perceptions and quotes that I felt exemplified the themes and supported my perceptions about raters and test-taker responses as recommended by Saldana (2016). There were 135 analyses (45 test-taker responses x 9 prompt/version combinations).

3.6.2.2 Second coding

In order to increase confidence in the codes and conclusions, the second step in the qualitative analysis was to conduct a round of second coding using the procedures outlined by Mackey and Gass (2015). It would have been difficult to share NVivo with my second coder so I transferred all the perceptions to Excel and constructed a 10% data set of rater perceptions, which is considered a minimum to establish confidence (Mackey & Gass, 2015, p 139). From the 4050 perceptions, I chose every 10th perception to be second coded for a total of 405 perceptions. I chose a pattern of perceptions which ensured an even number of perceptions for each rater and each test-taker response.

I next selected and trained a second coder. The second coder was a senior examiner who was part of the pilot studies so familiar with the goals of the study. I completed the training in two parts. The first part described the Excel coding instrument used in round 2 coding for themes (Appendix 15) and explained each code and the inclusion information. I provided the second coder with an information sheet with the following information: the research questions (section 2.4); the categories in the coding structure in round 2 (Appendix 15); the questions raters were asked to answer (Why and how did you make this on-topic decision?); a description of each code with inclusion information; and the tasks (Appendix 2). See Appendix 16 for the document sent to the second coder.

The second coder coded the first 15 perceptions in the sample by putting a ‘Y’ if the code was present in the rater perception (column B in Appendix 15). Follow-up training clarified difficulties, misunderstandings and differences.

The second coder coded the full 10% sample. I was available for support and conducted email exchanges with the second coder during and after the coding activity to collect issues and difficulties. I logged our email exchanges for reference in the analysis, with permission from the second coder.

I used exact inter-coder agreement, or simple percentage agreement because my data was continuous and not discrete.

3.6.2.3 Thematic analysis

The third analysis begins to answer RQ 1 and sub-research questions 1.1, 1.2, 1.3 and 1.4 (see section 2.4 for the research questions). I copied the rater perceptions into separate Excel files for each code for further analysis. I used the findings from the second coding to collapse (merge) codes and to conduct further quality assurance in my coding. I conducted frequency counts of key words used by raters in order to further understand the importance of each code (codes are in Appendix 15 and explained in section 3.6.2.1.2).

The analysis for the generic criteria codes included: amount of test-taker response; clarity of test-taker response; picture; words in prompt; examples; quotes; format (layout); pre-prepared test-taker response; no evidence; and off-topic decisions.

The analysis for task prompt type codes included: multiple activities, journeys or buildings (prompt 1); incorrect feelings (prompt 2); personal and general test-taker responses (prompt 3).

The analysis for the task version codes included: outdoor activities as a child (version 1); long journeys (version 2); and tall buildings (version 3).

The analysis for the rating scale codes included verb tenses, vocabulary and other/overall language.

3.6.2.4 Internal criteria raters use to make on-topic decisions

From the thematic analysis, I grouped findings and identified five internal criteria raters use to make on-topic decisions. Some codes were used more frequently and in a complimentary manner (e.g., the ‘quantity of response’ and ‘clarity of response’ are both related to ‘topic development).

The five categories that emerged from this analysis were: (1) Topic development; (2) question prompt content including grammar expectations, pictures and feelings; (3) vocabulary; (4) linguistic characteristics; and (5) approaches to collecting evidence.

3.6.2.5 Rater analysis

The fifth step in the qualitative analysis looked at each rater individually and documented evidence of severity, consistency and construct relevance in raters’ written perceptions. Rater severity looked for evidence of topic relevance perceptions that were either consistently very lenient or severe. In some cases, there was no pattern and they were inconsistently lenient and severe. Some raters had perceptions that were consistently not extreme (not very lenient or severe).

Rater consistency of perceptions considered their approach to making topic relevant decisions. I considered a rater to be consistent if they usually used the same decision-making process or vocabulary for each perception. A rater that I considered less consistent with their perceptions used different approaches to decide how and why a response was on-topic.

The final area I documented was the construct relevance of the raters' perceptions. Were their perceptions about speaking on-topic? Raters with low construct relevance discussed areas that I perceived were not about topic relevance. Raters with high construct relevance focused only on topic relevance.

I analysed perceptions for each rater across the three prompts and three versions and recorded my perceptions in Word. There were nine different analyses for each rater (three prompts x three versions) with 15 perceptions in each analysis. I collated these findings into an overall perception of each rater's severity, consistency and construct relevance judgements. See Table 7 for an example of this analysis.

From this analysis, I identified and collated raters who had very severe or very lenient perceptions. I also identified raters who had high and low rating consistency and construct relevance.

Table 7: Example of rater analysis

Rater 06 **P = Prompt 1, 2 or 3** **V = Version 1, 2 or 3**	
Overall	**Severity:** Seems consistently lenient across all prompts and versions (gives "benefit of the doubt", "it's close enough"). **Consistency:** Same short structure of perceptions for all versions in P1. Different types of perceptions in P2 and P3. **Construct relevance:** Focuses on main prompt words and tends to use an example/quote. Mentions clarity and detail. Some unusual perceptions for prompt 2. More variety of perceptions in prompt 3.
P1V1	- Focuses on prompt. Gives examples.
P1V2	- Uses trip and journey synonymously. Mentions 'long'.
P1V3	- Includes building name or location.
P2V1	- Interpretations of feelings: "Never really describes his emotions at the time except to say that he remembers it was a beautiful moment. But I'd give him the benefit of the doubt here." "Doesn't address question other than saying he loved to play cricket" ("beautiful moment" and "love" are not feelings).

	- Implied/implicit “Not much detail but says she ‘loved cycling’ and answers to Q1 & Q3 implies her enthusiasm”. - Examples/quotes – usually adjectives: “Felt good; it was fun.”
P2V2	- Seems easier to come up with feelings for this prompt “Says it was wonderful and she enjoyed it, then gives details of what exactly she enjoyed eg the food, picking up some of the language, the water sports, etc.” - Interesting perceptions “tell all the advantages and benefits of travel rather than specifically saying how she feels about it”. - Uses examples and some quotes.
P2V3	- Looks for examples and some quotes.
P3V1	- Impersonal: “Does not address the question at all. Only talks about himself and why he now as an adult, prefers calligraphy” - Reasons: “Addresses this fairly well; gives 2 clear reasons and attempts to give a third. Unfortunately, she becomes tongue-tied and stops well before time is up.”, “Gives a number/range of clear reasons” - Clear argument: “Clear, reasoned argument for benefits of outdoor activities”, “Logically argues from own experience that outdoor activities help children develop courage and confidence.”, “Somewhat stilted but clear argument for benefits of outdoor activities in childhood.” - Example/quote: “cricket prevents obesity! meeting people, get socialized” - Implied/implicit: “He doesn’t address this explicitly but discusses how it helped him learn discipline, spend time with family & friends, learn how to behave with others”.
P3V2	- Briefly: “Addresses briefly but clearly in the middle of her first response”, “Addresses briefly listing available means of transport” - Structure: “Compares present-day travel in her country (Egypt?) eg plane, bus, to historic, traditional travel eg camel, or boat. Coherent, well structured”. - Implied/implicit: “Unusual approach to answering — explains that Indonesians prefer spending money on shopping while traveling rather than mode of travel. Doesn’t specify, but implies that they choose cheapest mode possible in order to afford those purchases”. - Lists: “Lists different modes available and says which are used more commonly”. - Example/quote: “usually plane as most convenient; occasionally train”, “people lack economic resources to own car, so travel by bus, or several families travel together”.
P3V3	- Language: “doesn't give any reasons for building tall buildings but talks instead about why people like to visit (or live in?) tall buildings. Not quite the same thing but close enough perhaps in that it involves the same lexical families”, “says it’s important to have tall buildings and many people enjoy them; rather vague, but more a problem with language than relevance, I think.” - Benefit of the doubt: “shows countries’ development. In a way it’s borderline because he doesn’t say anything about the motive/intention of the people who build the tall buildings, rather draws conclusions of his own about why it’s a good thing. But for a speaking exam, it’s close enough!” - Example/quote: “the boss wanted to have a beautiful view, also have room for a lot of offices”

3.6.2.6 Qualitative analysis procedures conclusion

This section outlined the procedures used to conduct the qualitative analysis. See Table 8 for an overview of the qualitative data analysis steps.

Table 8: Qualitative data analysis procedures

Analysis step	Type of analysis	Purpose of analysis	Findings and products
Step 1A: Code raters' perceptions (round 1)	Open inductive coding using NVivo	To create initial codes and start to identify themes	Created initial codes Fine-tuned research questions
Step 1B: Coding rater perceptions (round 2)	Open inductive coding using NVivo with fine-tuned research questions	To confirm codes and themes	Confirmed codes and themes
Step 1C: Memos	Personal reflections written while coding	To document patterns or perception differences To take quotes to support my perceptions	135 analyses of perceptions (45 test-taker responses x 3 prompts)
Step 2: Second coding	Second code a 10% sample of test-taker responses using Excel Second coder used the codes from round 2	To determine that the coding was working as expected	Coding met the 75% threshold
Step 3: Thematic analysis	Recode perceptions using results from second coding Conduct frequency counts for codes and sub-codes	To quality assure the coding To quantify the data	Document findings for codes and sub-codes and provide evidence to answer research questions 1.1, 1.2, 1.3 and 1.4
Step 4: Identify internal criteria	Consolidate codes into categories	To bring together data into a smaller number of themes	Identify five internal criteria raters use to make on-topic decisions
Step 5: Rater analysis	Document evidence regarding the severity, consistency and construct relevance of each rater's perceptions	To understand the scope of rater perceptions with respect to severity, consistency and construct relevance	30 analyses of each raters' perceptions

3.6.3 Mixed methods interpretation procedures

The mixed methods interpretation integrated and triangulated the quantitative and qualitative data from the previous analyses as recommended by Bryman (2007), Caracelli & Greene (1993) and Creswell & Plano Clark (2018). The intent of this interpretation was to understand the relationship between severity, leniency, agreement and disagreement (fit statistics, unexpected decisions and bias) in the objective on-topic decisions and Aptis scores (quantitative data in section 4.1) with the thematic analysis and severity, consistency and construct relevance of the subjective perceptions (qualitative data in section 4.2) for a complete interpretation of the data. The findings are further triangulated with information from the email exchanges and missing data for a full response to RQ2: *What is the relationship between the internal criteria raters reportedly use to make their on-topic decisions and raters' Aptis scoring and on-topic decisions with respect to rating severity, reliability, consistency and construct relevance?*

I used a narrative discussion approach for integrating the quantitative and qualitative results, which Creswell and Plano Clark (2018) recommend for a convergent mixed methods study. I presented the quantitative results (Aptis and on-topic logits, fit statistics, unexpected observations, bias and missing data) and then the qualitative results for a selection of raters and test-taker results to exemplify key findings. The qualitative results are presented in the form of quotes from the rater perceptions and email exchanges and are used to confirm, disconfirm or complement the quantitative results.

There were two parts to the analysis. The first integration, triangulation and interpretation was to understand the relationship between the quantitative and

qualitative results for a selection of raters and the second part was for a selection of test-taker responses.

3.6.3.1 Raters

I first categorised the raters depending on their Aptis scores, on-topic decisions and on-topic perceptions into three groups. These groups are: (1) lenient raters; (2) severe raters; (3) and mixed profile raters. Lenient raters were raters that had Aptis scores and on-topic decisions with negative logits and lenient perceptions. Severe raters had the exact opposite profile (positive logits and severe perceptions). Mixed profile raters had lenient Aptis scores but severe on-topic decisions and/or perceptions (and vice versa). I selected raters who exemplified each group and wrote a narrative discussion. I documented the quantitative findings for each rater and then selected quotes from the perceptions and email exchanges that exemplified the findings.

I also include the raters' missing data and reported marking strategy when making topic relevance decisions from the email exchange to further complement the findings. I identified the extent to which the raters' rating process description matches the 2-step rating construct and highlighted key words and anything surprising as recommended by Bazeley (2013), Boyatzis (1998), Saldana (2016) and Wolcott (1994).

3.6.3.2 Test-taker responses

The second integration focused on the test-taker responses, again in a narrative approach. I examined test-taker responses with the highest, lowest and misfitting Aptis test-taker scores and explored possible explanations for these test-taker

scores by analysing rater on-topic decisions and on-topic perceptions. I also explored the lowest and misfitting on-topic decisions in a similar manner. The purpose of this analysis was to understand how rater on-topic decisions and perceptions affect Aptis scoring for a selection of test-taker responses.

4 Results

The data analysis and interpretation for the main study took place in two parts. The first part is the quantitative and qualitative analyses (sections 4.1 and 4.2), which I conducted separately and in parallel. The second part (section 4.3) is the mixed methods interpretation which integrates and triangulates the quantitative and qualitative findings into a deeper interpretation. See Table 9 for an overview of the procedures, including the data analysed, the analysis technique, the purpose of each analysis and the research question(s) addressed (see section 2.4 for the research questions). The next sections discuss these results in detail.

Table 9: Overview of data analysis

Analysis step	Data analysed	Data analysis technique	Purpose of analysis	Research questions
Quantitative analysis	Aptis scores and on-topic decisions	Many-Facet Rasch Measurement of on-topic decisions and Aptis scores using FACETS software	To understand rater severity, fit and bias for the Aptis scores To understand rater severity, fit and bias for the on-topic decisions	1.1 1.2 1.3
Qualitative analysis	On-topic written perceptions: how and why raters made on-topic decisions	Comparative thematic analysis of written perceptions using NVivo and Excel software	To understand on-topic decision patterns and themes for the sample of raters To identify severity, reliability and construct relevance of rater on-topic decisions	1 1.1 1.2 1.3 1.4
Mixed methods interpretation	All quantitative and qualitative data including email exchanges	Integration, interpretation and triangulation of data in a narrative discussion of raters and test-taker responses	To understand relationships between the quantitative and qualitative data To triangulate findings To validate results	2

4.1 Quantitative analysis results

I analysed the Aptis scores and on-topic decisions using Many-Facet Rasch Measurement (MFRM) and FACETS software (Linacre, 2018a).

The purpose was to understand rater severity, fit and bias for Aptis scores and on-topic decisions. I started by analysing the Aptis scores (45 per rater) and on-topic decisions (135 per rater) separately. I also investigated the prompts and versions. I kept the Aptis scores and on-topic decision analyses separate in order to understand the data, without initially influencing the findings by combining the data and the two different rating scales (7-point scale for Aptis scoring and 3-point scale for on-topic decisions). I also used the separate analyses in the mixed methods interpretation to compare with the on-topic perceptions. Finally, I combined all the data and the two rating scales into a bias analysis to investigate the interactions between the Aptis scores and on-topic decisions. This section explains these results.

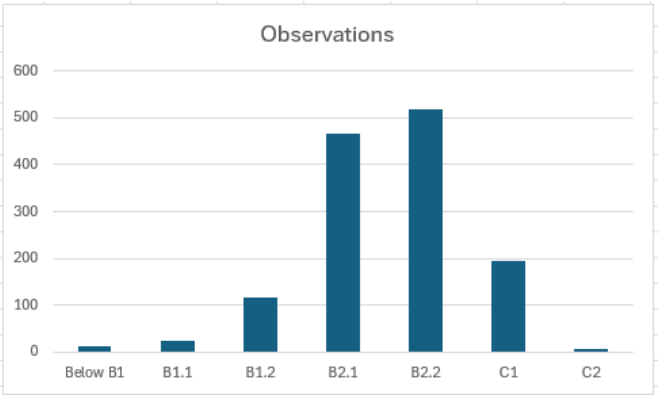
Before beginning the analyses, I confirmed that there were enough data points (observation counts) for the Aptis scores and on-topic decisions. As noted in section 3.5.1 (Quantitative data collection), Linacre (2002) and Eckes (2015, p. 132) recommend the sample size should have at least 30 observations per element and 10 observations in each rating scale category. The elements in this study, the test-taker responses (45) and the raters (30), meet the minimum of 30 observations per element.

There are two rating scale categories for Aptis scores that I am investigating (B2.1 and B2.2), which require a minimum of 20 observations (10 each). The three categories in the on-topic decisions (on-topic, borderline and off-topic) require a minimum of 30 observations, 10 in each category. The sample size for the Aptis

scores meets this minimum requirement (see Figure 4). There are 467 observations at the target B2.1 level and 519 observations at the B2.2 level. There were 1350 data points, therefore 364 responses were marked either below or above B2-level or were missing. The Aptis score observation count is a normal distribution about the B2 mean, median and mode.

Figure 4: Aptis score observation count

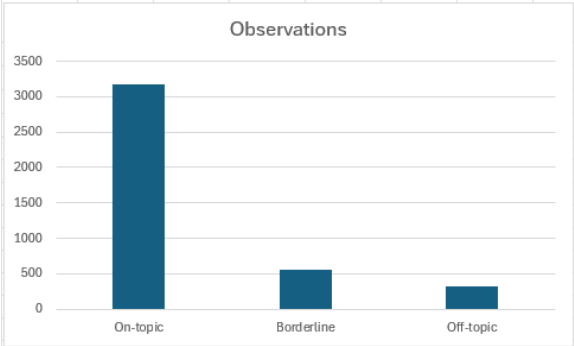
Aptis score category	Observations
Below B1	11
B1.1	24
B1.2	117
B2.1	467
B2.2	519
C1	194
C2	6
Missing	12



There are also a sufficient number of data points for the on-topic decisions (Figure 5). Of a total of 4050 observation points, there are 3181 on-topic, 547 borderline and 321 off-topic observations. There are fewer borderline and off-topic observations, as expected. At B2-level, most responses are on-topic and therefore it was a more realistic sample to have a large number of on-topic responses. The on-topic decisions are not a normal distribution with the mean, median and mode skewed in the on-topic category to the left.

Figure 5: On-topic observation count

On-topic decisions category	Observations
On-topic	3171
Borderline	547
Off-topic	321
Missing	11



There are at least 10 observations for each prompt and version categories (Tables 10 and 11). In the pilot studies where I collected the sample, I did not separate the rater decisions by prompt and version (see Section 3.4.3.4 Pilot study 3 findings for a discussion on this change for the main study) so I did not collect the test-taker response sample with an equal spread of topic relevance decisions with respect to prompts and versions. These observations are also not normal distributions, with the data skewed to the on-topic decisions.

Table 10: Observation count of topic relevance decisions for prompts

On-topic decisions category	Prompt 1 observations	Prompt 2 observations	Prompt 3 observations
On-topic	1158	1024	989
Borderline	136	223	188
Off-topic	55	98	168
Missing	1	5	5

Table 11: Observations count of topic relevance decisions for versions

On-topic decisions	Version 1 observations	Version 2 observations	Version 3 observations
On-topic	1040	1025	1106
Borderline	176	177	194
Off-topic	132	140	49
Missing	2	8	1

4.1.1 Aptis score analysis

The Aptis rating specification file is adapted from recommendations by Eckes (2015) and McNamara et al. (2019) and is in Figure 6. The specification file includes the rating scale model, rating scale, rater labels, test-taker response labels and data file, as well as instructions on the format of the output.

I used a 2-facet model (Models = ?,?,Aptis Scale). Facet 1 are the raters and facet 2 are the test-taker responses. The Aptis scores had a 7-point rating scale (<B1, B1.1, B1.2, B2.1, B2.2, C1 and C2 - see Appendix 4). Facet 2 (test-taker responses) is positive (Positive = 2) meaning that a higher measure represents higher Aptis scores. For raters, a higher measure indicates lower Aptis scores, or a rater who is more severe with Aptis scoring decisions, which is the convention in educational measurement (Eckes, 2015, pp. 56-57; Myford & Wolfe, 2003). I centred the raters around zero and non-centred the test-taker responses (Noncentered = 2). I set the vertical rulers to Vertical = 1A, 2A, so that the output shows individual rater and test-taker-response labels, rather than grouping them in any way. Agreement statistics were computed for raters (Inter-rater = 1). ‘Arrange = mN’ means that elements in each facet are arranged in the output tables according to their logit measures. Raters

were assigned a label of 1-30 (30 raters) and the test-taker responses were assigned labels 1-45 (45 test-taker responses).

Figure 6: Aptis scoring specification file

```
Title = Aptis Scores

Facets = 2
Positive = 2
Noncentered = 2
Vertical = 1A, 2A
Inter-rater = 1
Arrange = mN
*
Models = ?,?,Aptis Scale
*
Rating scale = Aptis Scale, R7
0 = 0_<B1
1 = 1_B1.1
2 = 2_B1.2
3 = 3_B2.1
4 = 4_B2.1
5 = 5_C1
6 = 6_C2
*
Labels =
1, Raters
1 = R06
2 = R07
3 = R08
4 = R09
5 = R10
6 = R11
7 = R12
8 = R13
9 = R14
10 = R15
11 = R16
12 = R17
13 = R18
14 = R19
15 = R20
16 = R21
17 = R22
18 = R23
19 = R24
20 = R25
21 = R26
22 = R27
23 = R28
24 = R29
25 = R30
26 = R31
27 = R33
28 = R34
29 = R35
30 = R36
*
```

```
2, Test-taker responses
1 = V1-107
2 = V1-112
3 = V1- 113
4 = V1-117
5 = V1-118
6 = V1-119
7 = V1-125
8 = V1-132
9 = V1-135
10 = V1-136
11 = V1-138
12 = V1-147
13 = V1-154
14 = V1-155
15 = V1-157
16 = V2-101
17 = V2-107
18 = V2-112
19 = V2-115
20 = V2-116
21 = V2-120
22 = V2-125
23 = V2-126
24 = V2-129
25 = V2-130
26 = V2-138
27 = V2-140
28 = V2-143
29 = V2-145
30 = V2-154
31 = V3-105
32 = V3-112
33 = V3-113
34 = V3-120
35 = V3-121
36 = V3-130
37 = V3-134
38 = V3-135
39 = V3-136
40 = V3-137
41 = V3-139
42 = V3-140
43 = V3-143
44 = V3-146
45 = V3-154
*
data = Aptis Scores data.xlsx
*
```

4.1.1.1 Aptis score: raters

The ‘score and measure’ table for raters in FACETS includes data advised by Eckes (2015, p. 71-82) and is in Table 12. I ordered the raters in column A by their severity measure (column B). Rater 13 (R13) is the most severe rater at the top with a logit measure of +1.43. Rater 08 (R08) is the most lenient at the bottom with a logit measure of -1.42.

The standard error (S.E) in column C shows the estimated degree of marking error by each rater. The standard error informs the researcher on confidence intervals, or how wide the possible marking could be for a rater (Eckes, 2015, pp. 72-73). For example, for rater 13, the most severe rater, the lower limit is 1.01 logits (1.43 - (2 x 0.21)) and the higher limit is 1.85 logits (1.43 + (2 x 0.21)). This rater overlaps with rater 07. The most lenient rater (R08), has a lower limit of -1.89 (-1.43 - (2 x 0.23)) and higher limit of -0.97 (-1.43 + (2 x 0.23)). Rater 08 overlaps with two other raters (R18 and R34). The most severe rater (R13) has a confidence interval of 0.84 logits (difference between lower and higher limits) while the most lenient rater has a confidence interval of 0.92 logits. The degree of uncertainty in the data is slightly higher for the most lenient rater than for the most severe.

For this research, I expected a restricted range of Aptis scores and for raters to mark all the responses in this sample at Aptis level 3 (B2.1) or 4 (B2.2) as the sample was purposefully chosen to be at B2-level (see Section 3.3 for a discussion on how the sample was selected to ensure the responses were at B2-level). The observed averages (mean average of Aptis scores) for each rater should therefore be in the 3s and 4s. All raters, except for rater 13, had observed averages in the 3s and 4s with 16 raters marking above 3.5 on average (more lenient) and 14 below 3.5 (more

severe). Rater 13 was the most severe rater with an observed average for Aptis scoring below 3 (i.e., B1-level). Fair average is also below 3.

Table 12: Raters: Aptis score and measure table

	A	B	C	D	E	F	G	H	I	J
1	Raters	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
2	R13	1.43	0.21	2.83	2.91	1.02	0.14	1.05	0.28	41
3	R07	1.11	0.2	3	3.06	0.93	-0.24	0.94	-0.21	45
4	R36	0.9	0.2	3.11	3.16	0.57	-2.23	0.56	-2.31	45
5	R09	0.82	0.21	3.17	3.2	0.94	-0.19	1.02	0.15	42
6	R33	0.82	0.21	3.16	3.2	1.47	1.92	1.5	2.04	45
7	R30	0.66	0.22	3.24	3.27	0.81	-0.79	0.83	-0.72	42
8	R15	0.65	0.21	3.24	3.28	0.51	-2.61	0.49	-2.82	45
9	R20	0.57	0.21	3.29	3.32	0.66	-1.66	0.65	-1.74	45
10	R16	0.52	0.21	3.31	3.34	0.69	-1.48	0.71	-1.42	45
11	R23	0.52	0.21	3.31	3.34	0.78	-1.03	0.79	-0.97	45
12	R11	0.35	0.21	3.4	3.42	1.31	1.36	1.36	1.55	45
13	R24	0.35	0.21	3.4	3.42	1.6	2.36	1.56	2.29	45
14	R12	0.3	0.21	3.42	3.44	0.59	-2.13	0.52	-2.62	45
15	R35	0.17	0.21	3.49	3.5	0.84	-0.68	0.87	-0.54	45
16	R17	0.12	0.21	3.51	3.52	1.27	1.19	1.23	1.08	45
17	R25	0.03	0.22	3.56	3.57	1.02	0.18	1.04	0.26	45
18	R14	-0.02	0.22	3.58	3.59	0.62	-1.96	0.64	-1.89	45
19	R26	-0.11	0.22	3.62	3.63	0.53	-2.54	0.54	-2.57	45
20	R21	-0.16	0.22	3.64	3.65	1.53	2.18	1.45	1.95	45
21	R22	-0.16	0.22	3.64	3.65	0.51	-2.7	0.52	-2.71	45
22	R10	-0.3	0.22	3.71	3.72	1.26	1.18	1.23	1.09	45
23	R19	-0.47	0.22	3.77	3.79	1.2	0.96	1.19	0.91	44
24	R27	-0.74	0.22	3.91	3.92	0.63	-1.98	0.62	-2.07	45
25	R28	-0.84	0.23	3.96	3.96	0.88	-0.54	0.86	-0.62	45
26	R31	-0.84	0.23	3.96	3.96	1.1	0.54	1.08	0.43	45
27	R29	-0.89	0.23	3.98	3.99	1.24	1.12	1.23	1.08	45
28	R06	-0.93	0.23	4	4	1.49	2.06	1.46	1.96	43
29	R34	-1.18	0.23	4.11	4.11	1.01	0.11	0.98	0	44
30	R18	-1.25	0.23	4.13	4.15	1.33	1.48	1.25	1.19	45
31	R08	-1.42	0.23	4.2	4.22	1.19	0.9	1.23	1.09	45

The observed average in column D is the mean average of the raw Aptis scores given by the rater. The ‘fair average’ in column E is the average expected rating for each rater and takes into consideration what the rater would normally score a test-taker response based on their pattern of decisions (Myford & Wolfe, 2004). The

observed and fair averages are similar, as expected, as all raters marked most responses (Eckes, 2015, p. 81-82). The most severe rater (R13) has an observed average Aptis score of 2.83 and a fair average of 2.91 so is on average marking below B2 level. I chose all responses to be at B2 level (scores of 3 and 4) so this rater is marking outside the expected CEFR level. There is a 1.31 score point fair average difference between the most severe and lenient raters (4.22 - 2.91 = 1.31), exceeding one Aptis band level.

The fit statistics are in columns F-I. The infit scores in column F exclude outliers, while the outfit scores in column H include the outliers. I have included both infit and outfit because outliers are expected and important in this study, due to the non-random sample of test-taker responses and a focus on responses that are difficult to mark. Columns G and I provide the standardised infit and outfit statistics and are used for significance testing. I have highlighted the InfitMS value above 1.3 to indicate rater misfit and an InfitZ value beyond +/- 2.0 indicates that the misfit is significant (McNamara, 1996). This InfitMS value threshold is similar to the upper limit of 1.365 recommended for this sample size (see Section 3.6.1 for a discussion on fit statistics and sample size). Eckes (2015) and Wolfe & McVay, 2012) recommend a higher outfit limit but as mentioned by Wright and Linacre (1994), this is a non-random sample so fit statistics can be idiosyncratic and should be used carefully, and I am expecting outliers. Outliers are key findings so I decided to also use an outfit level threshold of 1.3 for the investigation, if the outfits were significant. For Aptis scoring, raters 33, 11, 24, 21, 06 and 18 are misfitting and for raters 33, 24, 21 and 06, the misfits are significant.

Column J has the number of Aptis responses marked by raters. There were a total of 45 responses for 30 raters with 13 missing data for a total of 1337 data points. Raters 13, 06 and 34 entered two scores in the spreadsheet. Rater 30 entered “I’d escalate this”. Rater 09 put two scores for two decisions and “I’m not sure” for the third missing score. Rater 19 had no score in one Excel cell. These were all coded 9 in the data file (missing data). The effect of missing data on quantitative findings appears to be negligible; however, it does show that this rater might be ‘avoidant’ and ‘dependent’ raters as Baker (2012) identified. This point is raised further in the mixed methods interpretation when discussing particular raters.

I reviewed the rater separation and fixed chi-square statistics, which are used to understand how many severity levels are in the rater sample and if the finding is significant. Separation is the standard deviation divided by the average measurement error and estimates the number of statistically distinct levels of rater severity in the sample of raters (Eckes, 2015, pp. 62-63; Linacre 2018b, p. 327). It measures the spread of the rater severity relative to the precision of those measures (Myford & Wolfe, 2004, pp. 185-186). A rater separation index of 1.0 means that raters all have the same level of severity and are making the same decisions. Strata is a similar concept to separation but is used for data that is not normally distributed. The Fixed chi-square statistic is a significance test for the hypothesis that all raters in the sample are equally severe after accounting for measurement error (Eckes, 2015, p. 61). A significant chi-square value ($p < 0.01$) for a given sample of raters indicates that the severity estimates of at least two of the raters in the sample are statistically significantly different.

FACETS output includes separation, strata and chi-squared statistics. The population (Populn) model of the separation statistics is recommended for use in a fully crossed design (Eckes, 2015, p. 63). The separation statistic is recommended for a normal distribution and this is a normal distribution. The analysis found that there were 3.28 levels of Aptis scoring severity in the population statistics (see Figure 7). Raters can be separated into a little more than three statistically distinct levels of severity when scoring Aptis in this sample. This finding is higher than the two expected levels (B2.1 and B2.2). The Fixed chi-square value of 348.7 with 29 degrees of freedom (d.f.) and $p < 0.01$ is significant, which identifies that the raters are not equally severe in their Aptis scores. The exact agreement in Aptis decisions was 39.1% and the expected agreement was 36.3%.

Figure 7: Aptis separation and chi-squared statistics

```
Model, Populn: RMSE .21  Adj (True) S.D. .70  Separation 3.28  Strata 4.71  Reliability (not inter-rater) .91
Model, Sample: RMSE .21  Adj (True) S.D. .72  Separation 3.34  Strata 4.79  Reliability (not inter-rater) .92
Model, Fixed (all same) chi-squared: 348.7  d.f.: 29  significance (probability): .00
Model,  Random (normal) chi-squared: 26.8  d.f.: 28  significance (probability): .53
Inter-Rater agreement opportunities: 19200  Exact agreements: 7506 = 39.1%  Expected: 6975.7 = 36.3%
```

4.1.1.2 Aptis score: test-taker responses

I next reviewed the test-taker responses. The ‘score and measure’ table for the test-taker responses is in Table 13. Test-taker response V1-155 is the highest rated response. Responses V1-107 and V3-136 are jointly the lowest rated responses. V1-155, V1-119 and V2-145 have significant misfits. The test-taker responses with missing data are in column J.

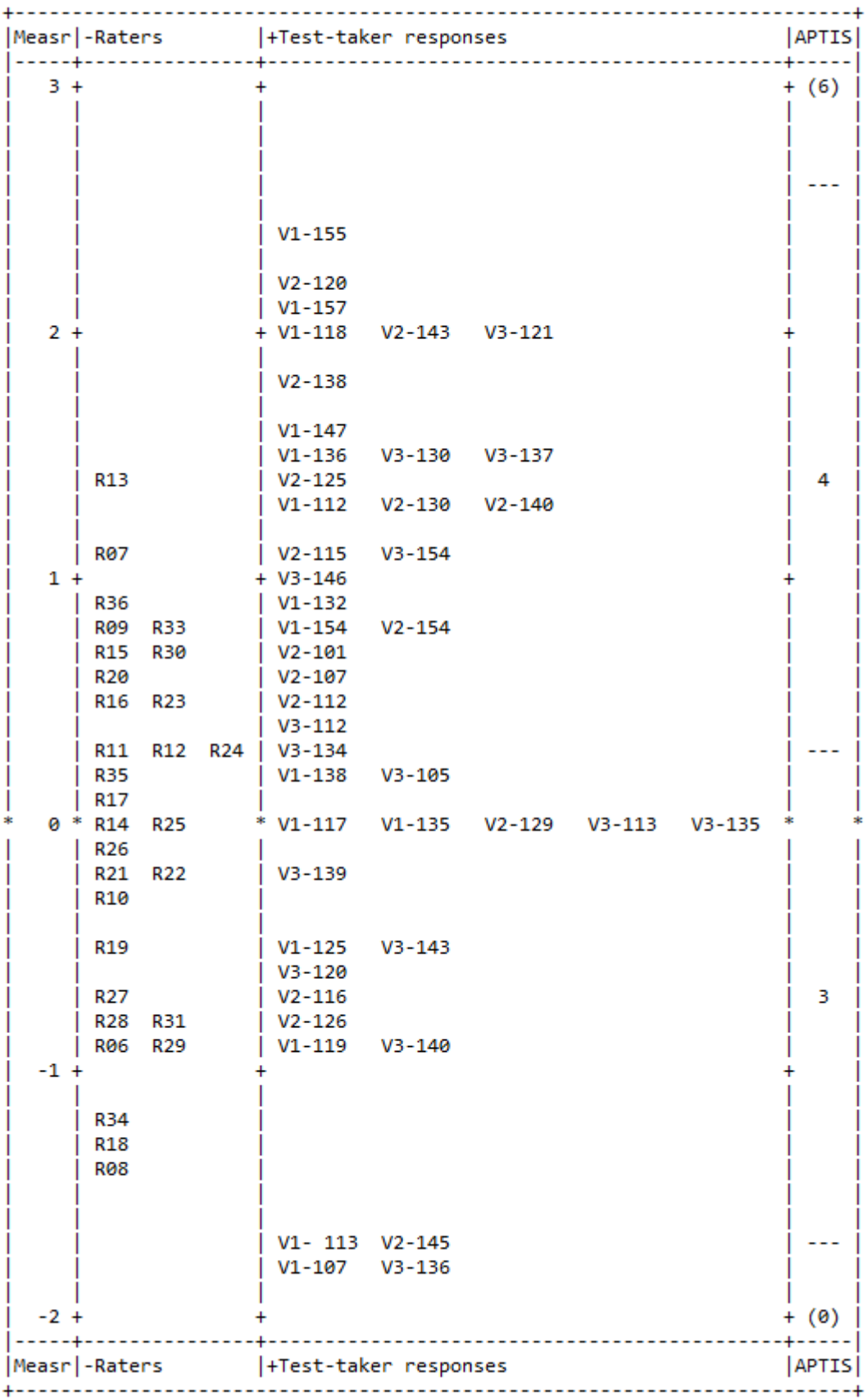
Table 13: Test-taker responses: Aptis score and measure table

	A	B	C	D	E	F	G	H	I	J
1	Test-taker	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
2	V1-155	2.38	0.3	4.41	4.41	1.78	2.55	1.87	2.76	29
3	V2-120	2.2	0.29	4.33	4.34	0.72	-1.17	0.69	-1.28	30
4	V1-157	2.12	0.28	4.3	4.3	0.95	-0.12	0.99	0.03	30
5	V1-118	1.99	0.29	4.24	4.25	0.77	-0.89	0.75	-1.01	29
6	V2-143	1.96	0.28	4.23	4.24	0.98	-0.01	0.94	-0.14	30
7	V3-121	1.96	0.28	4.23	4.24	0.65	-1.52	0.65	-1.52	30
8	V2-138	1.8	0.28	4.17	4.17	1.11	0.5	1.08	0.39	30
9	V1-147	1.64	0.28	4.1	4.1	0.72	-1.15	0.72	-1.15	30
10	V1-136	1.49	0.28	4.03	4.03	0.68	-1.35	0.68	-1.35	30
11	V3-130	1.49	0.28	4.03	4.03	0.59	-1.85	0.59	-1.86	30
12	V3-137	1.47	0.28	4.03	4.02	1.58	2.05	1.55	1.95	29
13	V2-125	1.42	0.28	4	4	1.14	0.62	1.12	0.56	30
14	V1-112	1.34	0.27	3.97	3.97	1.15	0.64	1.13	0.59	30
15	V2-130	1.34	0.27	3.97	3.97	0.8	-0.77	0.81	-0.75	30
16	V2-140	1.32	0.28	3.97	3.95	1.44	1.62	1.42	1.56	29
17	V2-115	1.12	0.27	3.87	3.86	0.95	-0.12	0.95	-0.13	30
18	V3-154	1.08	0.28	3.86	3.85	0.52	-2.2	0.53	-2.17	29
19	V3-146	1.04	0.27	3.83	3.83	0.87	-0.47	0.87	-0.47	30
20	V1-132	0.9	0.27	3.77	3.76	0.75	-0.97	0.75	-0.98	30
21	V1-154	0.75	0.27	3.7	3.7	0.99	0.05	0.99	0.04	30
22	V2-154	0.75	0.27	3.7	3.7	1.32	1.2	1.31	1.2	30
23	V2-101	0.68	0.27	3.67	3.66	1.11	0.48	1.12	0.54	30
24	V2-107	0.61	0.27	3.66	3.63	0.8	-0.71	0.81	-0.69	29
25	V2-112	0.46	0.27	3.57	3.56	1.53	1.82	1.54	1.86	30
26	V3-112	0.38	0.27	3.55	3.53	0.6	-1.68	0.59	-1.7	29
27	V3-134	0.32	0.27	3.5	3.5	0.89	-0.33	0.87	-0.44	30
28	V3-105	0.18	0.27	3.43	3.43	0.79	-0.76	0.79	-0.73	30
29	V1-138	0.17	0.27	3.41	3.43	0.96	-0.07	0.96	-0.05	29
30	V1-135	0.04	0.27	3.37	3.37	0.51	-2.09	0.52	-2.08	30
31	V3-113	0.04	0.27	3.37	3.37	1.03	0.19	1.03	0.21	30
32	V1-117	0	0.28	3.36	3.35	0.64	-1.35	0.64	-1.38	28
33	V2-129	-0.03	0.26	3.33	3.34	0.86	-0.45	0.87	-0.4	30
34	V3-135	-0.03	0.26	3.33	3.34	0.94	-0.11	0.97	-0.04	30
35	V3-139	-0.17	0.26	3.27	3.27	0.68	-1.21	0.68	-1.21	30
36	V3-143	-0.5	0.26	3.1	3.12	1.21	0.79	1.26	0.95	30
37	V1-125	-0.55	0.26	3.1	3.09	0.57	-1.68	0.58	-1.63	29
38	V3-120	-0.57	0.26	3.07	3.08	0.61	-1.52	0.6	-1.54	30
39	V2-116	-0.7	0.25	3	3.02	0.3	-3.36	0.29	-3.45	30
40	V2-126	-0.79	0.25	2.93	2.98	0.45	-2.33	0.43	-2.43	29
41	V3-140	-0.88	0.25	2.9	2.93	0.66	-1.24	0.69	-1.12	30
42	V1-119	-0.94	0.25	2.87	2.9	2.92	4.6	3.02	4.81	30
43	V2-145	-1.65	0.23	2.46	2.5	2.73	4.43	3.06	4.92	28
44	V1- 113	-1.7	0.22	2.4	2.47	0.63	-1.54	0.74	-0.98	30
45	V1-107	-1.79	0.22	2.33	2.41	1.13	0.57	1.03	0.2	30
46	V3-136	-1.79	0.22	2.33	2.41	0.6	-1.77	0.67	-1.32	30

Figure 8 is the Wright map, or variable map, which allows a visual comparison between the rater and test-taker response facets for Aptis scoring. It shows the logit measures (column 1), rater severity and leniency (column 2), the test taker responses (column 3) and the rating scale (column 4). Rater 13, the most severe rater, is at the top of column 2 and R08, the most lenient rater, is at the bottom of column. The response with the highest scores is at the top of column 3 (V1-155), while the responses with the lowest scores are at the bottom (V1-107 and V3-136).

The numbers in brackets at the top and bottom of column 4 are the rating scale maximum and minimum scores (6 and 0). The horizontal lines in column 4 are the Rasch-half-point thresholds. At the horizontal line point, a test-taker response would likely score half a point higher than the number below (e.g., 2.5) (McNamara et al., 2019, p. 100). Test-taker response V3-134 is at the 3.5 Rasch-half-point in this study. Test-taker responses below 3 have fair average scores below B2 level, identifying responses that, on average, are not being marked by raters at B2 level.

Figure 8: Aptis scores Wright map



The final report is the unexpected observations, which provides a list of the raters and test-taker responses that had very unexpected scores and is an additional check on if the model fits the data (Table 14). The observed scores (Obs) and expected scores (Exp) are transformed into standardised residuals (StRes).

In a rater reliability study using a random selection of test-taker responses, a satisfactory model fit is indicated when about 5% or fewer of (absolute) standardised residuals are >/= 2 and about 1% are or fewer of (absolute) standardised residuals are >/= 3 (Linacre 2018b, p. 169). These unexpected scores may be found in the output as misfits if they are severe and repetitive.

In this Aptis score data set, there are 1337 data points (3050 scores and 13 missing scores). There are 69 unexpected observations >/= 2 and eight unexpected observations >/= 3. The data has a satisfactory fit for unexpected observations >/= 3 (less than 13.37 observations or 1% of the total observations) but is slightly high for >/= 2 (more than 66.85 observations or 5%). This slightly high finding is expected as the test-taker responses were non-random and chosen to challenge raters' on-topic decisions and Aptis scoring and was not representative of normal marking.

The unexpected observations >/= 3 are in Table 14. Raters 08, 11, 19, 21 and 33 had unexpected Aptis scores. Rater 08 had two unexpected severe scores. Test-taker responses V1-155, V2-145, V3-137, V1-119, V2-145 and V1-119 had unexpected observations. Raters 11 and 33 had unexpected lenient Aptis scores (6=C2 and 5=C1) and raters 08, 19 and 21 had unexpected severe Aptis scores

Table 14: Aptis scores: unexpected observations

Raters	Test-taker responses	Obs	Exp	Res	Var	StRes
R08	V1-155	3	4.91	-1.91	0.3	-3.47
R08	V2-145	1	3.24	-2.24	0.46	-3.28
R11	V3-137	6	3.87	2.13	0.46	3.16
R19	V1-119	1	3.13	-2.13	0.47	-3.1
R21	V2-145	0	2.6	-2.6	0.6	-3.34
R33	V1-119	5	2.43	2.57	0.67	3.14

In one observation (Obs in Table 14) rater 21 scored zero (below B1 level) for test-taker response V2-145, while the expected score (Exp) is higher (2.6). In two observations, raters 08 and 19 scored 1 (B1.1 level). Keep in mind that all test-taker responses were at B2 level (3 or 4 score expected) so these raters had severe scores for these Aptis test-taker responses.

Two observations were unexpectedly high. Rater 11 scored 6 (C2-level) for V3-137 and rater 33 scored 5 (C1-level) for V1-119. For test-taker response V1-155, the highest scored response, rater 08 scored 3 (B2.1), which was severe. Test-taker response (V1-119) had a lenient score from rater 33 (5, C1 level) and a severe score from rater 19 (1, B1.1).

4.1.2 On-topic decision analysis (separate prompts)

I next analysed the on-topic decisions. I investigated the 135 on-topic decisions separately and I combined the prompts into the 45 test-taker responses. I analysed the decisions separately in order to understand on-topic decisions independently from Aptis scoring and full test-taker responses. I also combined the decisions for the three question prompts for each test-taker response so that I could compare the on-topic decisions with the Aptis scoring.

I first completed a 2-facet analysis for the 135 on-topic decisions by each rater (45 responses x 3 prompts). The specification file is in Figure 9 (Models = ?,?,On-topic scale). The two facets are the raters (facet 1) and the test-taker responses (facet 2). The rating scale is the on-topic rating scale with three categories (3 = On-topic; 2 = Borderline; 1 = Off-topic). Facet 2 is positive meaning that a higher measure represents more on-topic decisions. For raters (facet 1), a higher measure indicates a rater who has more off-topic and borderline decisions (more severe). I centred the raters (facet 1) around zero and non-centred the test-taker responses (Noncentred = 2). I set the vertical rulers to 1A, 2A to avoid grouping raters and test-taker responses on the Wright map. I ran agreement statistics for raters and the elements in each facet were arranged in the output according to their measures (mN). The rater numbers are the same as the Aptis scoring analysis (30 raters). There are 135 on-topic decisions labels (45 test-taker responses x 3 prompts).

Figure 9: On-topic decisions specification file (separate prompts)

```
Title = On-topic Decisions

Facets = 2
Positive = 2
Noncentered = 2
Vertical = 1A,2A
Inter-rater = 1
Arrange = mN
*
Models = ?,?,On-topic Scale
*
Rating scale = On-topic Scale, R3
3 = On-topic
2 = Borderline
1 = Off-topic
*
```

The model has the same rater labels (1-30). The 135 test-taker response labels include the version number (V1, V2, V3), the response number and the prompt (P1, P2, P3). An excerpt is in Figure 10.

Figure 10: On-topic test-taker response labels (separate prompts)

```
1 = V1-107-P1
2 = V1-112-P1
3 = V1-113-P1
…

16 = V2-101-P1
17 = V2-107-P1
18 = V2-112-P1
…

46 = V1-107-P2
47 = V1-112-P2
48 = V1-113-P2
…

133 = V3-143-P3
134 = V3-146-P3
135 = V3-154-P3
*
```

4.1.2.1 On-topic decisions: raters (separate prompts)

The ‘score and measure’ table displays the rater severity and fit measures for all on-topic decisions (Table 15). The column headings are the same as the Aptis scoring analysis. Rater 13 is the most severe on-topic rater and rater 28 is the most lenient. The observed and fair averages are in columns D and E. Rater 09 is misfitting (infit and outfit) and this finding is significant. Rater 29 has a significant outfit. Raters 10 and 08 are misfitting but the findings are not significant. There are missing data for raters 07, 21, 27 and 28 (T.Count less than 135), which are coded 9 and ignored in this analysis.

Table 15: Raters: on-topic score and measure table (separate prompts)

	A	B	C	D	E	F	G	H	I	J
1	Raters	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
2	R13	1.57	0.18	2.44	2.25	0.97	-0.13	0.89	-0.35	135
3	R34	1.04	0.2	2.55	2.48	1.13	0.84	1.23	0.91	135
4	R16	0.8	0.2	2.59	2.57	1.02	0.19	0.96	-0.05	135
5	R30	0.76	0.2	2.6	2.58	1.21	1.24	1.28	1.03	135
6	R33	0.76	0.2	2.6	2.58	1.23	1.34	1.29	1.08	135
7	R07	0.65	0.21	2.62	2.62	0.98	-0.09	0.89	-0.32	133
8	R09	0.64	0.21	2.62	2.62	1.37	2.03	1.76	2.33	135
9	R23	0.5	0.21	2.64	2.66	0.66	-2.19	0.62	-1.41	135
10	R25	0.5	0.21	2.64	2.66	0.81	-1.11	0.73	-0.91	135
11	R29	0.5	0.21	2.64	2.66	1.37	1.95	1.77	2.27	135
12	R31	0.46	0.22	2.65	2.67	0.75	-1.47	0.84	-0.49	135
13	R15	0.41	0.22	2.66	2.68	0.72	-1.71	0.65	-1.23	135
14	R24	0.22	0.23	2.69	2.73	1.1	0.6	0.81	-0.51	135
15	R17	0.17	0.23	2.7	2.74	0.76	-1.33	0.49	-1.82	135
16	R11	0.11	0.23	2.7	2.76	1.06	0.35	1.13	0.49	135
17	R26	0.06	0.23	2.71	2.77	0.9	-0.44	0.7	-0.84	135
18	R10	0.01	0.24	2.72	2.78	1.13	0.72	1.32	0.94	135
19	R22	-0.16	0.24	2.74	2.81	0.97	-0.05	0.58	-1.19	135
20	R21	-0.23	0.25	2.75	2.82	1	0.07	0.57	-1.15	129
21	R18	-0.55	0.27	2.79	2.87	1	0.08	0.85	-0.21	135
22	R36	-0.55	0.27	2.79	2.87	0.85	-0.63	0.6	-0.88	135
23	R12	-0.62	0.27	2.79	2.88	1.2	0.9	1	0.16	135
24	R20	-0.62	0.27	2.79	2.88	0.89	-0.41	0.48	-1.22	135
25	R14	-0.7	0.27	2.8	2.88	1.25	1.05	1.04	0.25	135
26	R19	-0.7	0.27	2.8	2.88	0.74	-1.12	0.46	-1.23	135
27	R06	-0.93	0.29	2.82	2.91	0.84	-0.61	0.89	-0.04	135
28	R08	-0.93	0.29	2.82	2.91	1.31	1.24	0.76	-0.29	135
29	R35	-0.93	0.29	2.82	2.91	1.09	0.45	1.16	0.46	135
30	R27	-1.07	0.3	2.83	2.92	1.15	0.63	1.13	0.41	133
31	R28	-1.19	0.31	2.84	2.93	0.68	-1.3	0.3	-1.39	134

The separation, strata and chi-squared statistics are in Figure 11. The decisions are clustered in the on-topic category and therefore are not normally distributed. Also, outliers are likely not accidental. Therefore, the strata statistic is more appropriate with 3.98 levels of on-topic decisions. The chi-squared is 263.3 with 29 degrees of

separation and is significant. The raters are making statistically different decisions. There is exact agreement in 77.2% of decisions with expected agreement of 77.4%.

Figure 11: On-topic strata and chi-squared statistics (separate prompts)

```
Model, Populn: RMSE .24  Adj (True) S.D. .66  Separation 2.73  Strata 3.98  Reliability (not inter-rater) .88
Model, Sample: RMSE .24  Adj (True) S.D. .68  Separation 2.79  Strata 4.05  Reliability (not inter-rater) .89
Model, Fixed (all same) chi-squared: 263.3  d.f.: 29  significance (probability): .00
Model,  Random (normal) chi-squared: 26.1  d.f.: 28  significance (probability): .57
Inter-Rater agreement opportunities: 58406  Exact agreements: 45068 = 77.2%  Expected: 45234.1 = 77.4%
```

4.1.2.2 On-topic decisions: test-taker responses (separate prompts)

I next analysed the test-taker responses for severity and fit. For the 135 on-topic decisions, there were 37 responses that were marked on-topic by all raters (Table 16). These responses have an Infit of 1 (perfect agreement). This is an expected finding as I wanted to include test-taker responses in the sample where all raters agreed that the response was on-topic. In addition, at B2 level most responses are on-topic so including a large number of on-topic responses in the sample was authentic. There were three missing data, which slightly affects the logits for those test-taker responses.

Table 16: Test-taker responses: perfect on-topic agreement (separate prompts)

	A	B	C	D	E	F	G	H	I	J
1	Test-taker responses	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
2	V1-112-P1	5.52	1.83	3	2.99	1	0	1	0	30
3	V1-112-P3	5.52	1.83	3	2.99	1	0	1	0	30
4	V1-113-P3	5.52	1.83	3	2.99	1	0	1	0	30
5	V1-117-P3	5.51	1.83	3	2.99	1	0	1	0	29
6	V1-118-P1	5.52	1.83	3	2.99	1	0	1	0	30
7	V1-118-P2	5.52	1.83	3	2.99	1	0	1	0	30
8	V1-119-P3	5.52	1.83	3	2.99	1	0	1	0	30
9	V1-132-P1	5.52	1.83	3	2.99	1	0	1	0	30
10	V1-135-P3	5.52	1.83	3	2.99	1	0	1	0	30
11	V1-138-P1	5.52	1.83	3	2.99	1	0	1	0	30
12	V1-154-P2	5.52	1.83	3	2.99	1	0	1	0	30
13	V1-157-P1	5.52	1.83	3	2.99	1	0	1	0	30
14	V1-157-P3	5.52	1.83	3	2.99	1	0	1	0	30
15	V2-101-P1	5.52	1.83	3	2.99	1	0	1	0	30
16	V2-116-P1	5.5	1.83	3	2.99	1	0	1	0	29
17	V2-120-P1	5.52	1.83	3	2.99	1	0	1	0	30
18	V2-120-P3	5.52	1.83	3	2.99	1	0	1	0	30
19	V2-130-P3	5.52	1.83	3	2.99	1	0	1	0	30
20	V2-138-P1	5.52	1.83	3	2.99	1	0	1	0	30
21	V2-138-P3	5.52	1.83	3	2.99	1	0	1	0	30
22	V2-140-P3	5.47	1.83	3	2.99	1	0	1	0	29
23	V2-143-P1	5.52	1.83	3	2.99	1	0	1	0	30
24	V2-143-P3	5.52	1.83	3	2.99	1	0	1	0	30
25	V3-112-P1	5.52	1.83	3	2.99	1	0	1	0	30
26	V3-112-P2	5.52	1.83	3	2.99	1	0	1	0	30
27	V3-113-P1	5.52	1.83	3	2.99	1	0	1	0	30
28	V3-120-P1	5.52	1.83	3	2.99	1	0	1	0	30
29	V3-121-P2	5.52	1.83	3	2.99	1	0	1	0	30
30	V3-130-P1	5.52	1.83	3	2.99	1	0	1	0	30
31	V3-130-P3	5.52	1.83	3	2.99	1	0	1	0	30
32	V3-135-P1	5.52	1.83	3	2.99	1	0	1	0	30
33	V3-135-P2	5.52	1.83	3	2.99	1	0	1	0	30
34	V3-137-P1	5.52	1.83	3	2.99	1	0	1	0	30
35	V3-139-P3	5.52	1.83	3	2.99	1	0	1	0	30
36	V3-143-P2	5.52	1.83	3	2.99	1	0	1	0	30
37	V3-146-P1	5.52	1.83	3	2.99	1	0	1	0	30
38	V3-154-P1	5.52	1.83	3	2.99	1	0	1	0	30

Table 17 has the test-taker responses with some on-topic disagreement but without any misfits. The fair averages are high so most examiners decided the responses were on-topic, but there were a few raters who marked the response either borderline or off-topic. These borderline or off-topic decisions were expected based

on the raters' on-topic scoring pattern, which explains why they are not identified as misfitting. There are two missing data highlighted.

Table 17: Test-taker responses: some on-topic disagreement (separate prompts)

	A	B	C	D	E	F	G	H	I	J
1	Test-taker responses	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
39	V1-155-P1	4.3	1.01	2.97	2.97	0.96	0.28	0.58	0.02	30
40	V2-107-P1	4.3	1.01	2.97	2.97	0.98	0.31	0.75	0.2	30
41	V2-115-P1	4.3	1.01	2.97	2.97	1.02	0.35	1.23	0.59	30
42	V2-125-P1	4.3	1.01	2.97	2.97	1.01	0.34	1.1	0.5	30
43	V3-105-P1	4.3	1.01	2.97	2.97	0.98	0.31	0.75	0.2	30
44	V3-143-P1	4.3	1.01	2.97	2.97	0.98	0.31	0.75	0.2	30
45	V2-101-P2	4.3	1.01	2.97	2.97	0.98	0.31	0.75	0.2	30
46	V2-120-P2	4.3	1.01	2.97	2.97	0.82	0.14	0.25	-0.46	30
47	V2-138-P2	4.3	1.01	2.97	2.97	0.92	0.25	0.44	-0.16	30
48	V3-137-P2	4.3	1.01	2.97	2.97	0.82	0.14	0.25	-0.46	30
49	V1-132-P3	4.3	1.01	2.97	2.97	0.96	0.28	0.58	0.02	30
50	V1-138-P3	4.3	1.01	2.97	2.97	0.97	0.3	0.65	0.1	30
51	V1-154-P3	4.3	1.01	2.97	2.97	0.97	0.3	0.65	0.1	30
52	V2-125-P3	4.3	1.01	2.97	2.97	0.97	0.3	0.65	0.1	30
53	V2-129-P3	4.3	1.01	2.97	2.97	0.92	0.25	0.44	-0.16	30
54	V3-113-P3	4.3	1.01	2.97	2.97	0.95	0.28	0.55	-0.01	30
55	V3-154-P3	4.3	1.01	2.97	2.97	0.92	0.25	0.44	-0.16	30
56	V2-116-P2	4.27	1.01	2.97	2.97	0.96	0.28	0.58	0.03	29
57	V1-135-P1	3.58	0.72	2.93	2.95	0.82	-0.05	0.5	-0.41	30
58	V1-136-P1	3.58	0.72	2.93	2.95	0.82	-0.05	0.52	-0.38	30
59	V1-147-P1	3.58	0.72	2.93	2.95	0.88	0.05	0.49	-0.43	30
60	V2-126-P1	3.58	0.72	2.93	2.95	1.06	0.31	2.02	1.2	30
61	V3-121-P1	3.58	0.72	2.93	2.95	0.88	0.05	0.49	-0.43	30
62	V3-140-P1	3.58	0.72	2.93	2.95	0.97	0.19	0.74	-0.03	30
63	V1-136-P2	3.58	0.72	2.93	2.95	0.97	0.18	1.26	0.56	30
64	V2-130-P2	3.58	0.72	2.93	2.95	1.03	0.28	1.17	0.47	30
65	V3-146-P2	3.58	0.72	2.93	2.95	0.75	-0.16	0.33	-0.76	30
66	V3-154-P2	3.58	0.72	2.93	2.95	1.04	0.29	1.91	1.12	30
67	V1-136-P3	3.15	0.6	2.9	2.92	0.98	0.15	0.83	-0.02	30
68	V3-146-P3	3.15	0.6	2.9	2.92	0.99	0.16	1.2	0.51	30
69	V2-107-P2	3.12	0.6	2.9	2.92	0.77	-0.26	0.49	-0.65	29
70	V1-154-P1	2.83	0.52	2.87	2.89	0.94	0.03	1.01	0.22	30
71	V2-140-P1	2.83	0.52	2.87	2.89	0.89	-0.09	0.66	-0.44	30
72	V1-147-P2	2.83	0.52	2.87	2.89	0.95	0.05	1.26	0.6	30

In Table 18 there are test-taker responses with more borderline or off-topic decisions and with misfits. Some raters are making unexpected topic relevance decisions. However, none of these misfits are significant. There are four missing data.

Table 18: Test-taker responses: more on-topic disagreement (separate prompts)

	A	B	C	D	E	F	G	H	I	J
1	Test-taker responses	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
73	V1-155-P2	2.83	0.5	2.87	2.89	1.51	1.07	1.62	1.07	30
74	V2-115-P2	2.83	0.5	2.87	2.89	0.95	0.05	0.78	-0.2	30
75	V2-143-P2	2.83	0.5	2.87	2.89	0.82	-0.23	0.54	-0.72	30
76	V3-136-P2	2.83	0.5	2.87	2.89	1.29	0.72	0.87	-0.03	30
77	V3-140-P2	2.83	0.5	2.87	2.89	0.82	-0.24	0.93	0.08	30
78	V3-121-P3	2.83	0.5	2.87	2.89	1.56	1.16	1.9	1.39	30
79	V2-140-P2	2.81	0.5	2.86	2.89	0.92	-0.01	0.72	-0.3	29
80	V1-147-P3	2.78	0.5	2.86	2.88	0.91	-0.02	0.67	-0.43	29
81	V1-117-P1	2.58	0.5	2.83	2.86	0.72	-0.58	0.52	-0.93	30
82	V1-118-P3	2.58	0.5	2.83	2.86	0.68	-0.69	0.5	-0.96	30
83	V2-154-P3	2.58	0.5	2.83	2.86	1.02	0.19	1.12	0.39	30
84	V2-126-P2	2.56	0.5	2.83	2.86	1.05	0.26	0.98	0.15	29
85	V3-134-P1	2.38	0.4	2.8	2.83	0.8	-0.42	0.97	0.1	30
86	V1-125-P3	2.38	0.4	2.8	2.83	0.81	-0.39	0.63	-0.73	30
87	V3-134-P3	2.38	0.4	2.8	2.83	1.13	0.46	0.76	-0.39	30
88	V3-143-P3	2.38	0.4	2.8	2.83	1.01	0.15	0.71	-0.51	30
89	V1-125-P2	2.2	0.4	2.77	2.8	0.89	-0.2	1.12	0.42	30
90	V1-132-P2	2.2	0.4	2.77	2.8	0.92	-0.12	0.9	-0.09	30
91	V2-125-P2	2.2	0.4	2.77	2.8	0.83	-0.4	0.92	-0.04	30
92	V3-120-P2	2.2	0.4	2.77	2.8	0.78	-0.54	0.7	-0.61	30
93	V3-112-P3	2.2	0.4	2.77	2.8	0.82	-0.42	0.95	0.03	30
94	V1-125-P1	2.04	0.4	2.73	2.77	0.69	-0.92	0.58	-1.08	30
95	V3-113-P2	2.02	0.4	2.72	2.77	0.74	-0.73	0.53	-1.29	29
96	V2-130-P1	1.89	0.4	2.7	2.74	1.15	0.56	1.01	0.16	30
97	V2-154-P1	1.89	0.4	2.7	2.74	1.27	0.89	1.31	0.87	30
98	V3-136-P1	1.89	0.4	2.7	2.74	1.27	0.89	1.55	1.36	30
99	V2-112-P2	1.89	0.4	2.7	2.74	1.26	0.86	1.27	0.78	30
100	V2-129-P2	1.89	0.4	2.7	2.74	0.7	-0.98	0.6	-1.12	30
101	V3-105-P2	1.89	0.4	2.7	2.74	0.92	-0.14	0.69	-0.78	30
102	V3-139-P2	1.89	0.4	2.7	2.74	0.76	-0.73	0.76	-0.55	30
103	V1-117-P2	1.76	0.4	2.67	2.71	1.37	1.21	1.19	0.63	30
104	V3-130-P2	1.76	0.4	2.67	2.71	0.96	-0.04	0.76	-0.62	30
105	V2-112-P1	1.63	0.3	2.63	2.68	0.94	-0.14	0.84	-0.39	30
106	V1-113-P1	1.52	0.3	2.6	2.64	0.53	-2.03	0.52	-1.72	30
107	V3-136-P3	1.52	0.3	2.6	2.64	1.12	0.53	1.26	0.86	30
108	V2-115-P3	1.41	0.3	2.57	2.61	1.31	1.15	1.3	0.99	30
109	V2-129-P1	1.31	0.3	2.53	2.58	0.92	-0.22	0.92	-0.19	30
110	V1-157-P2	1.31	0.3	2.53	2.58	1.05	0.27	1.13	0.52	30
111	V1-107-P1	1.21	0.3	2.5	2.54	1.05	0.27	1.12	0.5	30
112	V2-145-P3	1.21	0.3	2.5	2.54	1.18	0.76	1.11	0.47	30
113	V3-105-P3	1.21	0.3	2.5	2.54	0.94	-0.15	1.09	0.39	30
114	V1-107-P2	1.12	0.3	2.47	2.51	1.26	1.08	1.23	0.87	30
115	V2-154-P2	1.02	0.3	2.43	2.47	0.78	-0.94	0.78	-0.81	30

Table 19 has the most borderline and off-topic decisions as well as the most on-topic disagreement. One test-taker response (V1-112-P2) had a significant infit misfit.

Raters had significantly different perceptions on if this response was on-topic. V2-126-P3 and V1-119-P1 had significant outfit. V1-155-P3, V2-107-P3 and V2-145-P1 had misfits but the findings are not significant. There are two missing data.

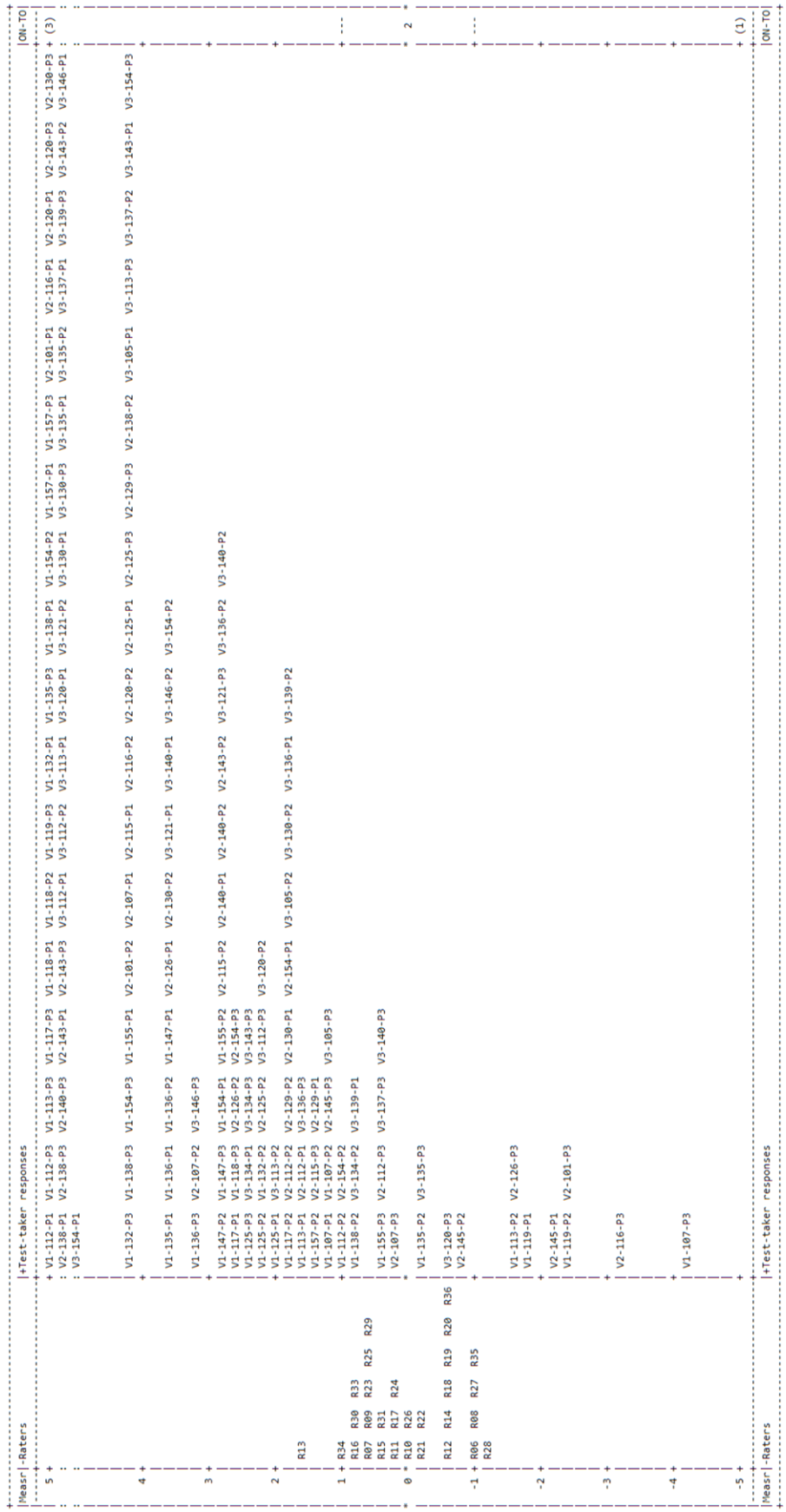
Table 19: Test-taker responses: most on-topic disagreement (separate prompts)

	A	B	C	D	E	F	G	H	I	J
1	Test-taker responses	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
116	V1-112-P2	0.94	0.3	2.4	2.44	1.62	2.39	1.67	2.32	30
117	V3-139-P1	0.85	0.3	2.37	2.4	0.76	-1.1	0.76	-0.99	30
118	V1-138-P2	0.85	0.3	2.37	2.4	0.65	-1.72	0.69	-1.34	30
119	V3-134-P2	0.77	0.3	2.33	2.37	0.92	-0.31	0.93	-0.23	30
120	V1-155-P3	0.45	0.3	2.2	2.22	1.26	1.23	1.31	1.39	30
121	V2-112-P3	0.37	0.3	2.17	2.19	0.97	-0.09	1.01	0.11	30
122	V3-137-P3	0.37	0.3	2.17	2.19	0.68	-1.72	0.67	-1.71	30
123	V3-140-P3	0.37	0.3	2.17	2.19	1.03	0.2	1.06	0.35	30
124	V2-107-P3	0.27	0.3	2.1	2.13	1.36	1.62	1.33	1.48	29
125	V1-135-P2	-0.15	0.3	1.93	1.93	0.98	-0.06	0.95	-0.17	30
126	V3-135-P3	-0.3	0.3	1.87	1.85	1.06	0.34	1.05	0.31	30
127	V3-120-P3	-0.53	0.3	1.77	1.74	1.04	0.24	0.99	0.04	30
128	V2-145-P2	-0.77	0.3	1.67	1.63	0.97	-0.08	0.97	-0.05	30
129	V1-113-P2	-1.52	0.3	1.4	1.36	0.96	-0.06	0.84	-0.42	30
130	V2-126-P3	-1.59	0.4	1.38	1.34	1.55	1.77	1.86	2.14	29
131	V1-119-P1	-1.76	0.4	1.33	1.29	1.68	2	1.86	2.04	30
132	V2-145-P1	-2.19	0.4	1.23	1.2	0.99	0.08	1.31	0.79	30
133	V1-119-P2	-2.37	0.4	1.2	1.17	0.84	-0.29	0.8	-0.28	30
134	V2-101-P3	-2.37	0.4	1.2	1.17	1.19	0.58	1.07	0.3	30
135	V2-116-P3	-3.14	0.6	1.1	1.08	0.82	-0.16	0.49	-0.66	30
136	V1-107-P3	-4.29	1	1.03	1.03	0.97	0.29	0.66	0.11	30

The Wright (variable) map for on-topic decisions is in Figure 12 and shows the logit measures (column 1), rater severity and leniency (column 2), the test taker responses (column 3) and the rating scale (column 4). The raters are centred around zero logits and the test-taker responses are floating.

Rater 13 is the most severe on-topic rater at the top of column 2 and rater 28 is the most lenient. Responses that are rated the most often on-topic are at the top in column 3 and the response rated the most often off-topic is at the bottom. Many more test-taker responses are scored on-topic, as expected, which explains why the data test-taker responses cluster at the top of column 3.

Figure 12: On-topic decisions Wright map (separate prompts)



I collated the 55 unexpected observations (>/= 3), which are in Table 20 and used in the mixed methods interpretation. What is notable is that almost all of the unexpected on-topic decisions are a rater marking a response borderline or off-topic.

Table 20: Unexpected on-topic observations (separate prompts)

Raters	Test-taker_responses	Obs	Exp	Res	StRes
R06	V3-154-P2	2	2.98	-0.98	-6.71
R06	V3-134-P1	2	2.93	-0.93	-3.58
R07	V3-143-P3	1	2.7	-1.7	-3.37
R07	V1-119-P1	3	1.16	1.84	4.73
R08	V2-126-P2	2	2.94	-0.94	-3.95
R09	V1-138-P3	2	2.95	-0.95	-4.32
R09	V1-154-P3	2	2.95	-0.95	-4.32
R09	V2-125-P3	2	2.95	-0.95	-4.32
R09	V3-136-P2	1	2.8	-1.8	-4.26
R09	V2-126-P3	3	1.19	1.81	4.32
R10	V2-115-P1	2	2.97	-0.97	-5.99
R10	V2-130-P2	2	2.95	-0.95	-4.12
R10	V2-112-P2	1	2.74	-1.74	-3.64
R11	V2-125-P1	2	2.97	-0.97	-5.67
R11	V1-136-P3	2	2.91	-0.91	-3.08
R12	V1-107-P1	1	2.73	-1.73	-3.54
R12	V1-107-P3	2	1.05	0.95	4.34
R14	V3-146-P3	2	2.96	-0.96	-4.75
R14	V3-140-P2	2	2.94	-0.94	-4.03
R14	V3-121-P3	2	2.94	-0.94	-4.03
R14	V1-112-P2	1	2.67	-1.67	-3.2
R15	V1-136-P1	2	2.92	-0.92	-3.31
R16	V3-113-P3	2	2.94	-0.94	-3.95
R16	V2-145-P1	2	1.09	0.91	3.03
R18	V1-136-P2	2	2.97	-0.97	-5.51
R18	V2-154-P3	2	2.92	-0.92	-3.26
R19	V1-154-P1	2	2.94	-0.94	-4.03
R22	V1-117-P2	1	2.75	-1.75	-3.71
R24	V2-126-P1	2	2.93	-0.93	-3.68
R24	V2-126-P3	3	1.28	1.72	3.5
R25	V2-101-P2	2	2.96	-0.96	-4.63
R26	V2-130-P2	2	2.94	-0.94	-4
R27	V1-147-P2	2	2.96	-0.96	-4.9
R27	V1-155-P2	2	2.96	-0.96	-4.9
R27	V2-125-P2	2	2.93	-0.93	-3.5
R29	V2-107-P1	2	2.96	-0.96	-4.63
R29	V3-105-P1	2	2.96	-0.96	-4.63
R29	V3-143-P1	2	2.96	-0.96	-4.63
R29	V3-140-P1	2	2.91	-0.91	-3.15
R30	V1-155-P1	2	2.94	-0.94	-4.04
R30	V1-132-P3	2	2.94	-0.94	-4.04
R30	V3-134-P3	1	2.67	-1.67	-3.17
R31	V1-135-P1	2	2.92	-0.92	-3.23
R31	V3-140-P1	2	2.92	-0.92	-3.23
R31	V3-154-P2	2	2.92	-0.92	-3.23
R33	V1-155-P2	1	2.78	-1.78	-3.99
R33	V2-116-P2	2	2.94	-0.94	-3.99
R33	V3-121-P3	1	2.78	-1.78	-3.99
R34	V2-138-P2	2	2.93	-0.93	-3.48
R34	V2-129-P3	2	2.93	-0.93	-3.48
R34	V3-154-P3	2	2.93	-0.93	-3.48
R35	V2-126-P1	2	2.98	-0.98	-6.71
R35	V3-121-P3	2	2.96	-0.96	-4.57
R35	V1-125-P2	2	2.92	-0.92	-3.25
R36	V3-136-P1	1	2.84	-1.84	-4.81

Having an analysis of the separate 135 on-topic decisions to understand rater on-topic decision severity, fit and unexpected on-topic decisions was useful to understand the data and for the mixed methods investigation. However, in the output for separate on-topic decisions, 1107 of the 4039 scores were considered 'responses in one extreme score', i.e., all raters made the same topic relevance decision. The software ignores these responses as these decisions do not provide additional information in the Rasch model. In addition, the three on-topic decisions for each test-taker response are combined for Aptis scoring decisions so I next combined the three on-topic decisions (P1, P2, P3) for each test-taker response.

4.1.3 On-topic decisions (combined prompts)

I conducted a 3-facet analysis (Models = ?,?,?,On-topic scale) so that the overall on-topic decisions for each response could be compared to the Aptis scores, and to remove the 'responses in one extreme score' issue mentioned in the previous section. Raters are expected to collate their three on-topic decisions for each test-taker response to fine-tune their Aptis score so it is important to understand leniency, fit and unexpected decisions for the full test-taker responses.

The specification file is in Figure 13. The three facets are the raters (facet 1), the test-taker responses (facet 2) and the prompts (facet 3). For each test-taker response there are on-topic decisions for three prompts (Dvalues = 3-1-3). For each facet 3 (prompts), there are three separate columns in the input data (1-3). The rating scale is the on-topic rating scale with three categories (3 = On-topic; 2 = Borderline; 1 = Off-topic). Facets 2 (test-taker responses) and 3 (prompts) are positive meaning that a higher measure represents a higher on-topic score. For raters, a higher measure indicates a rater who is more severe with topic relevance

decisions. I centred the raters (facet 1) around zero and non-centred facet 2 (test-taker responses) and facet 3 (prompts). I set the vertical rulers to 1A, 2A, 3A to avoid grouping. Inter-rater statistics and sorting (mN) remain the same as previous analyses. The rater and test-taker labels are the same as the Aptis scoring analysis, and there is an additional label for the prompts (facet 3).

Figure 13: On-topic decisions specification file (combined prompts)

```
Title = On-topic decisions (combined)

Facets = 3
Positive = 2,3
Noncentered = 2,3
Vertical = 1A,2A,3A
Inter-rater = 1
Arrange = mN
Dvalues = 3,1-3
*
Models = ?,?,?,On-topic Scale
*
Rating scale = On-topic Scale, R3
3 = On-topic
2 = Borderline
1 = Off-topic
*
```

The model has the same rater and test-taker labels as the Aptis analysis (30 raters and 45 test-taker responses). The third facet is the on-topic decisions by prompt:

```
*
3, On-topic decisions
1 = Prompt 1
2 = Prompt 2
3 = Prompt 3
*
data = On-topic (combined) data file.xlsx
*
```

An excerpt from the first few lines of the data file is in Table 21. The 30 raters (column A) mark the 45 test-taker responses (column B), and each test-taker response has three on-topic decisions (column C-E).

Table 21: Excerpt from the on-topic prompts data file for 3-facet analysis

	A	B	C	D	E
1	;Raters	Test-taker responses	On-topic Prompt 1	On-topic Prompt 2	On-topic Prompt 3
2	1	1	3	2	1
3	1	2	3	3	3
4	1	3	3	1	3
5	1	4	3	3	3
6	1	5	3	3	3
7	1	6	1	1	3
8	1	7	3	3	3
9	1	8	3	3	3
10	1	9	3	2	3
11	1	10	3	3	3

4.1.3.1 On-topic decisions: raters (combined prompts)

The ‘score and measure’ table in Table 22 displays the rating severity and fit measures for the on-topic decisions when collated for each test-taker response. The leniency and severity rank for raters does not change from the previous analysis (on-topic separate prompts). Rater 13 is still the most severe on-topic rater and rater 28 is the most lenient. However, taking the 1107 responses with full agreement into consideration in this analysis (which were ignored in the previous analysis) changes the fair averages and fit statistics. Rater 29 has an outfit (1.52), but it is not significant. Rater 34 has an outfit of 1.6 and it is significant (2.04). Raters 07, 21, 27 and 28 have missing data.

Table 22: Raters: on-topic score and measure table (combined prompts)

	A	B	C	D	E	F	G	H	I	J
1	Raters	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
2	R13	0.84	0.13	2.44	2.59	1.01	0.13	1.21	0.94	135
3	R34	0.55	0.14	2.55	2.7	1.16	1.18	1.6	2.04	135
4	R16	0.42	0.15	2.59	2.74	0.95	-0.32	0.97	-0.02	135
5	R30	0.4	0.15	2.6	2.75	1.06	0.46	1.18	0.69	135
6	R33	0.4	0.15	2.6	2.75	1.04	0.29	1.16	0.62	135
7	R07	0.34	0.15	2.62	2.76	1.14	0.97	1.08	0.35	133
8	R09	0.33	0.15	2.62	2.76	1.03	0.28	1.01	0.14	135
9	R23	0.26	0.16	2.64	2.78	0.75	-1.77	0.6	-1.43	135
10	R25	0.26	0.16	2.64	2.78	1.11	0.76	0.92	-0.16	135
11	R29	0.26	0.16	2.64	2.78	1.15	0.97	1.52	1.57	135
12	R31	0.24	0.16	2.65	2.79	0.96	-0.24	0.96	-0.02	135
13	R15	0.21	0.16	2.66	2.79	0.83	-1.1	0.82	-0.52	135
14	R24	0.11	0.16	2.69	2.82	0.98	-0.05	0.76	-0.7	135
15	R17	0.08	0.17	2.7	2.82	0.87	-0.79	0.61	-1.26	135
16	R11	0.05	0.17	2.7	2.83	1.09	0.57	1.03	0.19	135
17	R26	0.03	0.17	2.71	2.83	0.97	-0.1	0.73	-0.74	135
18	R10	0	0.17	2.72	2.84	1	0.04	0.86	-0.3	135
19	R22	-0.09	0.18	2.74	2.85	1.26	1.38	1	0.12	135
20	R21	-0.1	0.18	2.75	2.86	1.03	0.23	0.65	-0.92	129
21	R18	-0.29	0.19	2.79	2.88	1.25	1.25	0.88	-0.17	135
22	R36	-0.29	0.19	2.79	2.88	1.02	0.17	0.7	-0.69	135
23	R12	-0.33	0.19	2.79	2.89	0.78	-1.09	0.73	-0.58	135
24	R20	-0.33	0.19	2.79	2.89	1.01	0.11	0.56	-1.14	135
25	R14	-0.36	0.19	2.8	2.89	1.11	0.59	0.88	-0.17	135
26	R19	-0.36	0.19	2.8	2.89	1.04	0.28	0.8	-0.38	135
27	R06	-0.48	0.2	2.82	2.91	1.08	0.42	0.85	-0.22	135
28	R08	-0.48	0.2	2.82	2.91	0.95	-0.17	0.57	-1.01	135
29	R35	-0.48	0.2	2.82	2.91	1.07	0.38	0.73	-0.52	135
30	R27	-0.56	0.21	2.83	2.91	0.88	-0.46	0.65	-0.7	133
31	R28	-0.62	0.22	2.84	2.92	1.11	0.54	0.48	-1.22	134

The strata statistics found 2.83 levels of severity for the combined on-topic decisions (Figure 14). Raters were grouped into almost three levels of severity. The chi-squared statistic of 139.8 with 29 degrees of freedom is significant (.00), therefore identifying that raters are not making topic relevance decisions with the same level of severity. Exact agreement is 77.2% with expected agreement of 69.7%.

Figure 14: On-topic strata and chi-square (combined prompts)

```
Model, Populn: RMSE .17  Adj (True) S.D. .33  Separation 1.87  Strata 2.83  Reliability (not inter-rater) .78
Model, Sample: RMSE .17  Adj (True) S.D. .33  Separation 1.91  Strata 2.88  Reliability (not inter-rater) .78
Model, Fixed (all same) chi-squared: 139.8  d.f.: 29  significance (probability): .00
Model,  Random (normal) chi-squared: 23.9  d.f.: 28  significance (probability): .69
Inter-Rater agreement opportunities: 58406  Exact agreements: 45068 = 77.2%  Expected: 40690.6 = 69.7%
```

4.1.3.2 On-topic decisions: test-taker responses (combined prompts)

The on-topic ‘score and measure’ table for the test-taker full responses is in Table 23. Test-taker responses V2-120 and V2-138 have the highest number of on-topic decisions while V2-145 has the lowest number of on-topic decisions. Test-taker responses V1-119 and V2-145 have significant misfits. V1-154, V2-130 and V1-112 have insignificant misfits. Missing data is highlighted in column J and discussed in the mixed methods interpretation.

Table 23: Test-taker responses: score and measure table (combined prompts)

	A	B	C	D	E	F	G	H	I	J
1	Test-taker responses	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
2	V2-120	4.47	0.99	2.99	2.99	0.94	0.3	0.41	-0.26	90
3	V2-138	4.47	0.99	2.99	2.99	0.95	0.31	0.56	-0.05	90
4	V3-154	3.41	0.56	2.97	2.97	0.9	0.06	0.91	0.11	90
5	V2-143	3.14	0.48	2.96	2.96	0.87	-0.05	0.59	-0.53	90
6	V1-118	2.93	0.43	2.94	2.95	0.77	-0.3	0.4	-1.15	90
7	V1-154	2.93	0.43	2.94	2.95	0.9	-0.02	1.33	0.7	90
8	V3-146	2.93	0.43	2.94	2.95	0.81	-0.21	0.55	-0.74	90
9	V3-121	2.76	0.39	2.93	2.94	1.16	0.47	0.99	0.16	90
10	V1-136	2.62	0.36	2.92	2.93	0.82	-0.28	0.75	-0.38	90
11	V3-112	2.62	0.36	2.92	2.93	0.76	-0.43	0.57	-0.85	90
12	V3-143	2.62	0.36	2.92	2.93	0.99	0.13	0.63	-0.67	90
13	V1-132	2.5	0.34	2.91	2.92	0.81	-0.35	0.71	-0.51	90
14	V2-140	2.47	0.34	2.91	2.92	0.85	-0.24	0.92	-0.01	88
15	V2-125	2.39	0.32	2.9	2.91	0.8	-0.42	0.79	-0.36	90
16	V3-113	2.39	0.32	2.9	2.91	0.91	-0.09	0.55	-1.05	89
17	V3-130	2.3	0.3	2.89	2.91	1.09	0.36	0.76	-0.48	90
18	V1-147	2.28	0.3	2.89	2.9	0.76	-0.57	0.68	-0.71	89
19	V2-130	2.21	0.29	2.88	2.9	1.06	0.29	1.58	1.35	90
20	V1-157	1.99	0.25	2.84	2.87	1	0.11	0.97	0.03	90
21	V1-117	1.92	0.25	2.83	2.86	1.1	0.43	1	0.11	89
22	V1-112	1.76	0.23	2.8	2.83	1.33	1.22	1.33	1.03	90
23	V2-115	1.76	0.23	2.8	2.83	0.91	-0.25	0.73	-0.81	90
24	V1-138	1.67	0.21	2.78	2.81	0.77	-0.92	0.71	-0.98	90
25	V1-125	1.62	0.21	2.77	2.8	0.69	-1.4	0.74	-0.88	90
26	V2-129	1.5	0.2	2.73	2.77	0.85	-0.64	1.12	0.52	90
27	V3-105	1.46	0.19	2.72	2.76	0.71	-1.46	0.59	-1.74	90
28	V3-136	1.46	0.19	2.72	2.76	0.9	-0.43	1.14	0.59	90
29	V3-137	1.42	0.19	2.71	2.75	0.69	-1.58	0.53	-2.11	90
30	V3-139	1.35	0.18	2.69	2.73	0.91	-0.41	1.21	0.86	90
31	V1-155	1.32	0.18	2.68	2.72	0.91	-0.4	0.79	-0.87	90
32	V2-107	1.27	0.18	2.66	2.7	0.97	-0.07	0.77	-0.97	88
33	V2-154	1.26	0.18	2.66	2.7	0.88	-0.62	0.94	-0.18	90
34	V3-140	1.26	0.18	2.66	2.7	0.77	-1.31	0.68	-1.5	90
35	V3-134	1.23	0.17	2.64	2.69	0.83	-0.96	0.79	-0.92	90
36	V1-135	1.17	0.17	2.62	2.67	1.17	1	1	0.06	90
37	V3-135	1.17	0.17	2.62	2.67	0.94	-0.28	0.79	-0.97	90
38	V3-120	0.9	0.15	2.51	2.56	0.79	-1.58	0.7	-1.8	90
39	V2-112	0.88	0.15	2.5	2.55	0.78	-1.64	0.76	-1.45	90
40	V2-101	0.66	0.14	2.39	2.43	1.14	1.17	1.03	0.25	90
41	V2-126	0.65	0.15	2.39	2.43	1.06	0.53	0.97	-0.12	88
42	V2-116	0.58	0.14	2.34	2.38	1.14	1.22	1.04	0.3	88
43	V1-113	0.56	0.14	2.33	2.37	1.24	2.02	1.17	1.2	90
44	V1-107	0	0.13	2	2	0.95	-0.55	0.92	-0.72	90
45	V1-119	-0.25	0.14	1.84	1.83	1.94	7.51	2.17	7.99	90
46	V2-145	-0.32	0.14	1.8	1.78	1.45	3.92	1.52	3.95	90

Table 24 shows the ‘score and measure’ table for the on-topic prompts using the combined prompt data. Prompt 1 on-topic decisions have the highest measure (i.e., responses are more likely to be marked on-topic by raters). Prompt 3 is the lowest with fewer on-topic decisions, with prompt 2 in the middle. There are no misfitting

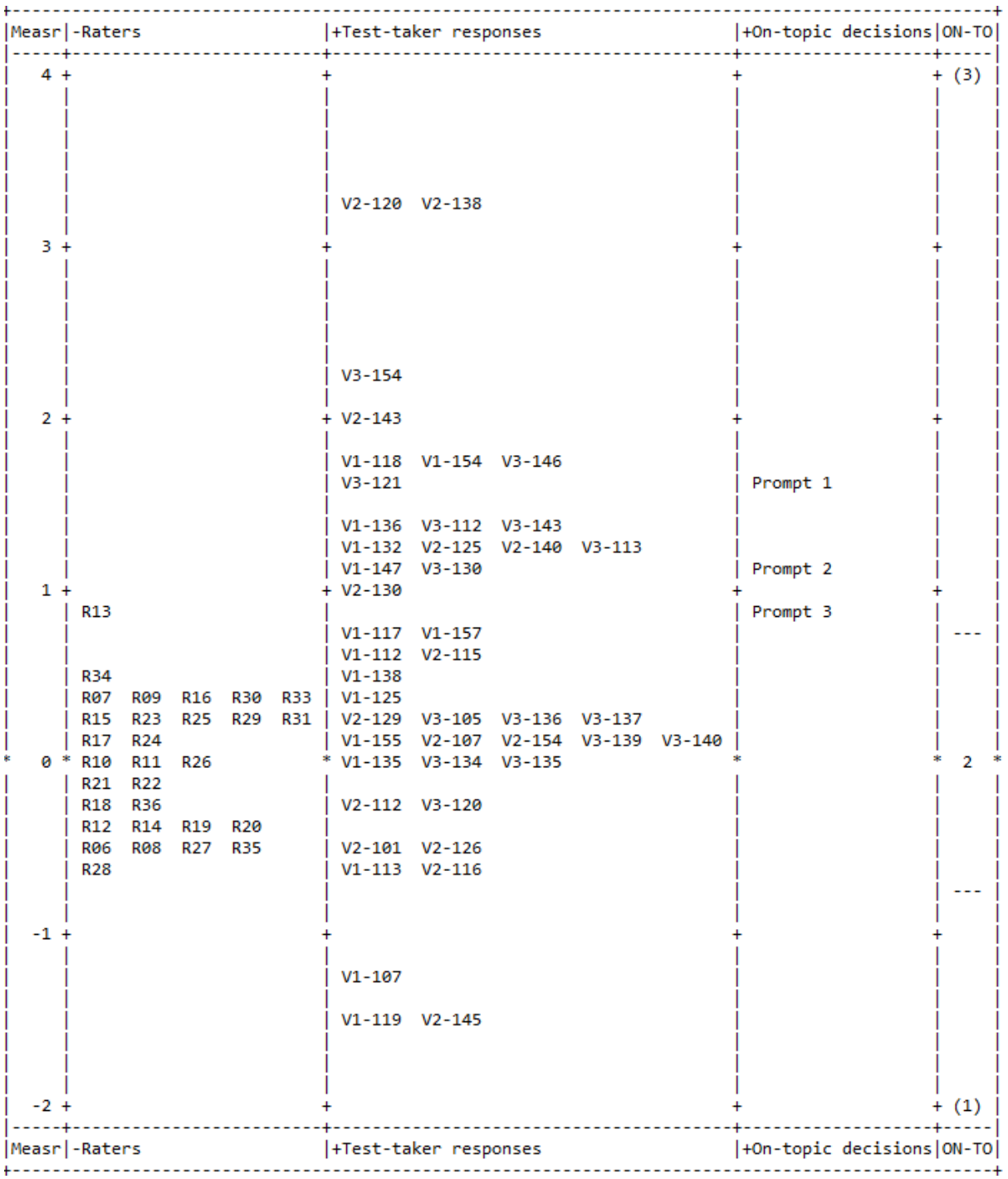
prompts. Prompt 1 has one missing score, prompts 2 and 3 each have five missing scores (1350 data points expected for each prompt). As mentioned earlier, the sample was not chosen to have an equal number of topic relevant decisions for each category (on-topic, borderline, off-topic) so it is not possible to determine if the observed and fair average rank is important.

Table 24: Prompts: on-topic score and measure table (combined prompts)

	A	B	C	D	E	F	G	H	I	J
1	On-topic decisions	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
2	Prompt 1	0.47	0.06	2.82	2.9	0.85	-2.31	0.75	-1.95	1349
3	Prompt 2	-0.1	0.05	2.69	2.82	0.98	-0.46	0.99	-0.03	1345
4	Prompt 3	-0.36	0.05	2.61	2.76	1.16	3.32	0.91	-0.96	1345

The Wright map for on-topic decisions for combined prompts is in Figure 15 and shows the logit measures (column 1), rater severity and leniency (column 2), the test-taker responses (column 3), the order of the prompts (column 4) and the rating scale (column 5). Rater 13 is the most severe rater at the top of column 2 and rater 28 is the most lenient at the bottom. Responses that are rated most on-topic are at the top in column 3 (V2-120 and V2-138) and the response rated most off-topic is at the bottom (V2-145). Similar to the separate on-topic decisions in the previous analysis, most responses were rated on-topic and clustered at the top of column 3. In column 4, prompt 1 at the top is rated most often on-topic and prompt 3 is rated most often off-topic, although as noted above, it is not possible to determine if this finding is important. The rating scale (1-3) is in column 5.

Figure 15: On-topic decisions Wright map (combined prompts)



There are 64 unexpected observations >/= 3. This is slightly higher than the unexpected observations for separate prompts, which was 55, and many of the observations are different, implying that there is a difference between on-topic decisions in the individual prompts, and when combining the prompts for each response. For example, a rater could be harsh with prompt 1 for a particular test taker, but then lenient with prompts 2 and 3, balancing out the on-topic decisions when a full test-taker response is considered.

Table 25: Unexpected on-topic observations (combined prompts)

Raters	Test-taker responses	On-topic decisions	Obs	Exp	Res	StRes
R06	V3-154	Prompt 2	2	2.98	-0.98	-6.93
R07	V3-143	Prompt 3	1	2.86	-1.86	-4.62
R07	V2-143	Prompt 2	2	2.94	-0.94	-3.64
R08	V1-155	Prompt 3	1	2.75	-1.75	-3.35
R08	V3-136	Prompt 1	2	2.92	-0.92	-3.03
R09	V1-117	Prompt 2	1	2.77	-1.77	-3.45
R09	V2-140	Prompt 1	2	2.93	-0.93	-3.45
R10	V2-130	Prompt 1	2	2.94	-0.94	-3.62
R10	V2-115	Prompt 3	1	2.74	-1.74	-3.26
R11	V2-140	Prompt 1	2	2.95	-0.95	-4.11
R11	V2-125	Prompt 1	2	2.95	-0.95	-3.91
R12	V2-130	Prompt 1	2	2.96	-0.96	-4.42
R12	V3-112	Prompt 3	2	2.93	-0.93	-3.4
R13	V2-120	Prompt 2	2	2.97	-0.97	-5.97
R13	V1-119	Prompt 3	3	1.24	1.76	3.38
R13	V3-113	Prompt 2	1	2.75	-1.75	-3.35
R13	V3-130	Prompt 2	1	2.73	-1.73	-3.14
R14	V1-112	Prompt 2	1	2.87	-1.87	-4.96
R14	V3-146	Prompt 3	2	2.95	-0.95	-4.23
R14	V3-121	Prompt 3	2	2.94	-0.94	-3.81
R15	V1-136	Prompt 1	2	2.95	-0.95	-4.09
R16	V2-143	Prompt 2	2	2.93	-0.93	-3.46
R16	V2-140	Prompt 1	2	2.93	-0.93	-3.25
R16	V1-117	Prompt 2	1	2.74	-1.74	-3.24
R17	V1-118	Prompt 3	2	2.92	-0.92	-3.18
R18	V1-112	Prompt 2	1	2.86	-1.86	-4.73
R18	V1-136	Prompt 2	2	2.95	-0.95	-3.92
R18	V1-132	Prompt 2	2	2.94	-0.94	-3.63
R18	V1-135	Prompt 2	1	2.73	-1.73	-3.15
R19	V1-154	Prompt 1	2	2.98	-0.98	-6.83
R20	V3-112	Prompt 3	2	2.93	-0.93	-3.4
R21	V1-132	Prompt 2	2	2.92	-0.92	-3.21
R22	V1-157	Prompt 2	1	2.87	-1.87	-4.82
R22	V1-117	Prompt 2	1	2.85	-1.85	-4.6

R22	V1-112	Prompt 2	1	2.83	-1.83	-4.14
R24	V2-129	Prompt 1	1	2.85	-1.85	-4.44
R24	V2-130	Prompt 1	2	2.93	-0.93	-3.37
R25	V3-130	Prompt 2	1	2.86	-1.86	-4.67
R25	V1-112	Prompt 2	1	2.74	-1.74	-3.25
R26	V2-130	Prompt 1	2	2.94	-0.94	-3.55
R27	V2-125	Prompt 2	2	2.95	-0.95	-4.01
R27	V1-147	Prompt 2	2	2.94	-0.94	-3.73
R28	V1-157	Prompt 2	2	2.92	-0.92	-3.22
R29	V2-130	Prompt 1	1	2.92	-1.92	-6.36
R29	V1-154	Prompt 1	2	2.96	-0.96	-4.78
R29	V3-143	Prompt 1	2	2.95	-0.95	-3.96
R29	V2-143	Prompt 2	2	2.94	-0.94	-3.83
R29	V2-154	Prompt 1	1	2.76	-1.76	-3.39
R30	V1-154	Prompt 1	2	2.96	-0.96	-4.4
R30	V1-157	Prompt 2	1	2.77	-1.77	-3.46
R30	V2-140	Prompt 1	2	2.93	-0.93	-3.3
R31	V3-154	Prompt 2	2	2.96	-0.96	-4.59
R31	V2-130	Prompt 1	2	2.92	-0.92	-3.09
R33	V3-121	Prompt 3	1	2.87	-1.87	-4.87
R33	V1-154	Prompt 1	2	2.96	-0.96	-4.4
R33	V3-121	Prompt 1	2	2.95	-0.95	-3.97
R33	V3-139	Prompt 1	1	2.75	-1.75	-3.3
R34	V2-138	Prompt 2	2	2.98	-0.98	-7.01
R34	V3-121	Prompt 1	2	2.94	-0.94	-3.61
R34	V3-154	Prompt 3	2	2.92	-0.92	-3.2
R34	V2-143	Prompt 2	2	2.92	-0.92	-3.18
R35	V3-121	Prompt 3	2	2.95	-0.95	-4.1
R35	V3-140	Prompt 3	1	2.74	-1.74	-3.2
R35	V2-129	Prompt 1	2	2.92	-0.92	-3.11
R36	V3-136	Prompt 1	1	2.9	-1.9	-5.62

4.1.4 On-topic decision analysis by prompt

I next did separate 2-facet analyses for the the three prompts and three versions. I split the test-taker responses in the specification file by only including either prompt 1, 2, or 3 or version 1, 2 or 3 input data. Rater and test-taker response labels are the same as previous analyses.

Table 26 shows the score and measure table for prompt 1. Rater 29 is the most severe with a logit measure of 2.37. Rater 14 is the most lenient with a logit measure

of -2.01. Raters 33, 10, 11, 12, 34, 09, 16, 07, 08, 20 are misfitting but these findings are not significant. Eckes (2015, pp. 78-81) notes that sample size affects the standardised fit statistics in that rater marking deviations in larger sample sizes lead to more pronounced significance. As expected, with a smaller sample of only 15 test-taker responses for prompts and versions, significance is more difficult to achieve.

Table 26: Raters: on-topic score and measure table (prompt 1)

	A	B	C	D	E	F	G	H	I	J
1	Raters	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
2	R29	2.37	0.37	2.56	2.26	1.21	0.82	1.2	0.59	45
3	R13	1.63	0.41	2.67	2.51	0.83	-0.51	0.73	-0.56	45
4	R31	1.63	0.41	2.67	2.51	0.64	-1.28	0.62	-0.92	45
5	R30	1.46	0.42	2.69	2.56	0.96	-0.03	1.08	0.33	45
6	R24	1.1	0.44	2.73	2.66	0.84	-0.43	0.69	-0.55	45
7	R33	0.89	0.46	2.76	2.71	1.63	1.69	1.67	1.25	45
8	R10	0.68	0.48	2.78	2.76	1.02	0.18	1.5	0.95	45
9	R11	0.68	0.48	2.78	2.76	0.93	-0.1	1.41	0.82	45
10	R26	0.68	0.48	2.78	2.76	0.78	-0.57	0.7	-0.39	45
11	R12	0.44	0.5	2.8	2.8	1.5	1.31	1	0.21	45
12	R23	0.44	0.5	2.8	2.8	0.63	-1.07	0.43	-0.92	45
13	R25	0.44	0.5	2.8	2.8	0.67	-0.91	0.4	-1.02	45
14	R34	0.44	0.5	2.8	2.8	1.11	0.41	1.46	0.84	45
15	R15	0.18	0.53	2.82	2.84	0.97	0.06	1.1	0.38	45
16	R36	0.18	0.53	2.82	2.84	1.17	0.55	0.7	-0.21	45
17	R09	-0.12	0.56	2.84	2.88	1.52	1.21	1.36	0.67	45
18	R16	-0.12	0.56	2.84	2.88	1.39	0.98	1.23	0.53	45
19	R17	-0.12	0.56	2.84	2.88	0.78	-0.44	0.5	-0.44	45
20	R35	-0.12	0.56	2.84	2.88	0.7	-0.65	1.14	0.44	45
21	R21	-0.46	0.61	2.86	2.91	0.5	-1.18	0.23	-0.8	44
22	R06	-0.87	0.67	2.89	2.94	0.71	-0.46	0.64	0.08	45
23	R07	-0.87	0.67	2.89	2.94	1.94	1.6	0.79	0.23	45
24	R19	-0.87	0.67	2.89	2.94	0.73	-0.4	1	0.41	45
25	R22	-0.87	0.67	2.89	2.94	0.47	-1.13	0.18	-0.64	45
26	R08	-1.37	0.75	2.91	2.96	1.99	1.5	0.84	0.43	45
27	R18	-1.37	0.75	2.91	2.96	0.48	-0.91	0.13	-0.4	45
28	R20	-1.37	0.75	2.91	2.96	1.89	1.4	0.56	0.2	45
29	R27	-1.37	0.75	2.91	2.96	0.95	0.12	0.6	0.24	45
30	R28	-1.37	0.75	2.91	2.96	0.48	-0.91	0.13	-0.4	45
31	R14	-2.01	0.85	2.93	2.98	0.69	-0.26	0.12	0.04	45

The unexpected observations for prompt 1 are in Table 27. Except for rater 26 and 33, all of these raters had on-topic decisions that were too low for these test-taker responses. Both raters decided that test-taker response (V2-145) was borderline (Obs = 2) while the expected decision off-topic (1.1 and 1.08), i.e., most raters perceived the response to be off-topic (Exp = 1).

Table 27: Unexpected observations (prompt 1)

Raters	Test-taker responses	Obs	Exp	Res	Var	StRes
R06	V3-134	2	2.93	-0.93	0.06	-3.71
R08	V3-136	2	2.93	-0.93	0.07	-3.58
R09	V2-140	2	2.92	-0.92	0.08	-3.3
R10	V2-115	2	2.96	-0.96	0.04	-4.92
R11	V2-125	2	2.96	-0.96	0.04	-4.92
R15	V1-136	2	2.95	-0.95	0.05	-4.31
R16	V2-140	2	2.92	-0.92	0.08	-3.3
R19	V1-154	2	2.96	-0.96	0.04	-4.85
R26	V2-145	2	1.1	0.9	0.09	3.01
R27	V2-154	2	2.93	-0.93	0.07	-3.58
R30	V1-155	2	2.92	-0.92	0.08	-3.29
R33	V2-145	2	1.08	0.92	0.07	3.37
R34	V1-147	2	2.94	-0.94	0.06	-3.77
R34	V3-121	2	2.94	-0.94	0.06	-3.77
R35	V2-126	2	2.96	-0.96	0.04	-5.01
R36	V3-136	1	2.72	-1.72	0.22	-3.69

Table 28 shows the score and measure table for prompt 2. Rater 13 is the most severe with a logit measure of 2.18. Rater 35 is the most lenient with a logit measure of -2. Raters 33, 22, 10, 26, 29, 06, 14 are misfitting but these are not significant.

Table 28: Raters: on-topic score and measure table (prompt 2)

	A	B	C	D	E	F	G	H	I	J
1	Raters	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
2	R13	2.18	0.28	2.22	2.08	1.07	0.39	1.02	0.17	45
3	R16	1.55	0.29	2.4	2.37	0.84	-0.67	0.72	-0.86	45
4	R34	1.46	0.29	2.42	2.4	1.27	1.17	1.31	1	45
5	R09	1.1	0.31	2.51	2.54	1.22	0.93	1.12	0.45	45
6	R15	1.1	0.31	2.51	2.54	0.56	-2.08	0.49	-1.71	45
7	R23	1	0.32	2.53	2.57	0.71	-1.2	0.63	-1.08	45
8	R07	0.69	0.33	2.6	2.66	0.86	-0.46	0.77	-0.47	45
9	R25	0.69	0.33	2.6	2.66	1.13	0.54	1.25	0.72	45
10	R30	0.69	0.33	2.6	2.66	1.08	0.38	0.85	-0.24	45
11	R33	0.58	0.34	2.62	2.69	1.36	1.28	1.9	1.83	45
12	R22	0.33	0.36	2.67	2.75	1.31	1.07	0.82	-0.25	45
13	R11	0.2	0.37	2.69	2.78	0.68	-1.11	0.63	-0.68	45
14	R17	0.06	0.38	2.71	2.8	0.51	-1.81	0.38	-1.35	45
15	R18	0.06	0.38	2.71	2.8	0.99	0.09	1.01	0.19	45
16	R27	0.06	0.38	2.71	2.8	0.77	-0.69	0.83	-0.16	45
17	R31	0.06	0.38	2.71	2.8	0.71	-0.94	1.16	0.47	45
18	R10	-0.09	0.4	2.73	2.83	1.58	1.63	1.83	1.35	45
19	R24	-0.25	0.41	2.76	2.85	1.08	0.34	0.58	-0.6	45
20	R26	-0.43	0.43	2.78	2.87	1.16	0.55	1.6	0.98	45
21	R29	-0.43	0.43	2.78	2.87	1.25	0.76	1.38	0.73	45
22	R21	-0.59	0.45	2.78	2.89	0.81	-0.41	0.53	-0.51	41
23	R06	-0.62	0.45	2.8	2.89	0.87	-0.24	1.32	0.63	45
24	R12	-0.74	0.48	2.84	2.91	0.76	-0.49	0.55	-0.39	44
25	R28	-1.05	0.5	2.84	2.93	0.83	-0.28	0.35	-0.62	44
26	R14	-1.07	0.5	2.84	2.93	1.82	1.7	1.44	0.73	45
27	R19	-1.07	0.5	2.84	2.93	0.61	-0.94	0.22	-0.96	45
28	R20	-1.07	0.5	2.84	2.93	0.71	-0.62	0.49	-0.37	45
29	R36	-1.07	0.5	2.84	2.93	0.93	-0.02	0.65	-0.11	45
30	R08	-1.34	0.53	2.87	2.95	1.1	0.37	1.01	0.36	45
31	R35	-2	0.62	2.91	2.97	1.15	0.44	1.21	0.63	45

Unexpected observations are in Table 29. All these decisions are lower than expected.

Table 29: Unexpected observations: prompt 2

	A	B	C	D	E	F	G	H
1	Raters	Test-taker responses	Obs	Exp		Res	Var	StRes
2	R06	V3-154	2	2.98	0.98	-0.98	0.02	-6.3
3	R08	V2-126	2	2.97	0.97	-0.97	0.03	-5.28
4	R09	V3-136	1	2.74	1.74	-1.74	0.22	-3.73
5	R10	V2-130	2	2.96	0.96	-0.96	0.04	-4.8
6	R10	V2-112	1	2.78	1.78	-1.78	0.19	-4.08
7	R10	V2-115	2	2.91	0.91	-0.91	0.08	-3.19
8	R12	V1-125	2	2.91	0.91	-0.91	0.08	-3.13
9	R14	V3-140	2	2.97	0.97	-0.97	0.03	-5.32
10	R14	V1-112	1	2.77	1.77	-1.77	0.2	-3.93
11	R18	V1-136	2	2.95	0.95	-0.95	0.05	-4.44
12	R20	V3-139	2	2.91	0.91	-0.91	0.08	-3.13
13	R22	V1-117	1	2.65	1.65	-1.65	0.28	-3.1
14	R25	V2-101	2	2.96	0.96	-0.96	0.04	-4.68
15	R26	V2-130	2	2.97	0.97	-0.97	0.03	-5.71
16	R29	V2-115	2	2.94	0.94	-0.94	0.06	-3.82
17	R29	V2-143	2	2.94	0.94	-0.94	0.06	-3.82
18	R31	V3-154	2	2.95	0.95	-0.95	0.05	-4.44
19	R33	V2-116	2	2.96	0.96	-0.96	0.04	-4.93
20	R33	V1-155	1	2.84	1.84	-1.84	0.15	-4.81
21	R34	V2-138	2	2.91	0.91	-0.91	0.09	-3.11
22	R35	V1-125	2	2.97	0.97	-0.97	0.03	-6.04
23	R36	V3-120	2	2.93	0.93	-0.93	0.06	-3.73

Table 30 shows the score and measure table for prompt 3. Rater 07 is the most severe and rater 27 is the most lenient. Raters 09, 30, 24, 21, 11, 08 and 12 are misfitting. Raters 9 has a significant infit misfit and 30 has a significant outfit misfit.

Table 30: Raters: on-topic score and measure table (prompt 3)

	A	B	C	D	E	F	G	H	I	J
1	Raters	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
2	R07	1.45	0.34	2.35	2.08	0.84	-0.54	0.58	-0.57	43
3	R13	1.17	0.34	2.42	2.22	1.03	0.21	0.68	-0.43	45
4	R33	1.17	0.34	2.42	2.22	1.06	0.32	0.84	-0.08	45
5	R34	1.17	0.34	2.42	2.22	1.02	0.17	1.26	0.61	45
6	R09	0.69	0.35	2.51	2.44	1.71	2.15	3.01	2.53	45
7	R30	0.69	0.35	2.51	2.44	1.74	2.22	1.89	1.44	45
8	R16	0.57	0.36	2.53	2.49	0.86	-0.42	1.26	0.61	45
9	R17	0.57	0.36	2.53	2.49	1.01	0.13	0.64	-0.53	45
10	R25	0.57	0.36	2.53	2.49	0.63	-1.4	0.45	-1.03	45
11	R24	0.3	0.37	2.58	2.59	1.34	1.13	1.04	0.27	45
12	R26	0.3	0.37	2.58	2.59	0.67	-1.15	0.37	-1.19	45
13	R31	0.3	0.37	2.58	2.59	0.75	-0.82	0.61	-0.56	45
14	R21	0.27	0.38	2.61	2.6	1.34	1.12	0.81	-0.12	44
15	R23	0.17	0.38	2.6	2.63	0.53	-1.74	0.77	-0.19	45
16	R29	0.17	0.38	2.6	2.63	0.99	0.08	1.11	0.37	45
17	R14	0.02	0.39	2.62	2.67	1.13	0.52	1.05	0.3	45
18	R20	0.02	0.39	2.62	2.67	0.66	-1.14	0.42	-0.94	45
19	R10	-0.13	0.39	2.64	2.71	0.86	-0.33	0.63	-0.38	45
20	R11	-0.13	0.39	2.64	2.71	1.58	1.63	1.43	0.78	45
21	R15	-0.13	0.39	2.64	2.71	0.52	-1.72	0.65	-0.35	45
22	R19	-0.29	0.4	2.67	2.75	0.92	-0.15	0.45	-0.71	45
23	R22	-0.29	0.4	2.67	2.75	0.75	-0.71	0.36	-0.94	45
24	R08	-0.45	0.42	2.69	2.79	1.35	1.03	0.93	0.18	45
25	R36	-0.45	0.42	2.69	2.79	0.55	-1.46	0.36	-0.85	45
26	R35	-0.63	0.43	2.71	2.82	1.15	0.51	1	0.28	45
27	R18	-0.82	0.45	2.73	2.85	1	0.12	0.81	0.09	45
28	R12	-1.03	0.46	2.76	2.87	1.3	0.83	1.46	0.73	45
29	R06	-1.25	0.49	2.78	2.9	0.87	-0.18	0.44	-0.26	45
30	R28	-1.25	0.49	2.78	2.9	0.62	-0.91	0.35	-0.4	45
31	R27	-2.76	0.64	2.88	2.98	0.76	-0.29	0.24	0.36	43

The unexpected observations for prompt 3 are in Table 31. In all three instances, the on-topic decision is higher than expected.

Table 31: Unexpected observations (prompt 3)

Raters	Test-taker responses	Obs	Exp	Res	Var	StRes
R09	V2-126	3	1.17	1.83	0.16	4.6
R12	V1-107	2	1.06	0.94	0.06	3.93
R24	V2-126	3	1.24	1.76	0.22	3.76

4.1.5 On-topic decision analysis by version

Table 32 shows the score and measure table for version 1. Rater 13 is the most severe. Raters 08 and 35 are jointly the most lenient. Raters 30, 09, 07, 33, 12, 29, 14, 20 and 08 are misfitting but these findings are not significant

Table 32: Raters: on-topic score and measure table (version 1)

	A	B	C	D	E	F	G	H	I	J
1	Raters	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
2	R13	1.61	0.33	2.42	2.15	0.83	-0.62	0.74	-0.25	45
3	R30	1.5	0.33	2.44	2.21	0.94	-0.15	1.5	0.92	45
4	R26	1.15	0.35	2.51	2.36	0.84	-0.51	0.57	-0.73	45
5	R09	1.03	0.35	2.53	2.41	1.4	1.35	1.95	1.55	45
6	R15	1.03	0.35	2.53	2.41	0.69	-1.13	0.69	-0.45	45
7	R16	1.03	0.35	2.53	2.41	1.11	0.46	0.73	-0.37	45
8	R34	0.91	0.36	2.56	2.47	1.05	0.26	1.03	0.24	45
9	R23	0.77	0.37	2.58	2.52	0.75	-0.81	0.61	-0.66	45
10	R31	0.77	0.37	2.58	2.52	0.8	-0.61	0.86	-0.09	45
11	R25	0.64	0.38	2.6	2.56	0.8	-0.58	0.57	-0.73	45
12	R07	0.52	0.39	2.61	2.6	1.26	0.87	1.42	0.84	44
13	R17	0.49	0.39	2.62	2.61	0.52	-1.74	0.38	-1.26	45
14	R22	0.49	0.39	2.62	2.61	1.18	0.65	0.7	-0.4	45
15	R11	0.34	0.4	2.64	2.66	0.84	-0.41	0.83	-0.12	45
16	R33	0.34	0.4	2.64	2.66	1.58	1.62	1.64	1.12	45
17	R18	0.18	0.41	2.67	2.7	1.04	0.23	1.03	0.25	45
18	R24	0.18	0.41	2.67	2.7	0.89	-0.21	0.54	-0.71	45
19	R21	-0.18	0.44	2.71	2.78	0.98	0.08	0.54	-0.58	45
20	R27	-0.18	0.44	2.7	2.78	0.7	-0.76	0.9	0.08	44
21	R12	-0.38	0.46	2.73	2.82	1.71	1.62	1.99	1.28	45
22	R06	-0.6	0.48	2.76	2.85	0.44	-1.62	0.23	-1.14	45
23	R29	-0.6	0.48	2.76	2.85	1.26	0.72	1.52	0.82	45
24	R19	-0.84	0.5	2.78	2.88	0.58	-1.02	0.86	0.14	45
25	R36	-0.84	0.5	2.78	2.88	0.76	-0.45	0.4	-0.55	45
26	R14	-1.1	0.53	2.8	2.91	1.88	1.71	0.89	0.24	45
27	R20	-1.1	0.53	2.8	2.91	1.35	0.85	0.58	-0.13	45
28	R28	-1.38	0.55	2.82	2.93	0.62	-0.8	0.34	-0.36	45
29	R10	-1.7	0.58	2.84	2.95	0.6	-0.82	0.33	-0.19	45
30	R08	-2.05	0.61	2.87	2.96	1.7	1.32	0.92	0.51	45
31	R35	-2.05	0.61	2.87	2.96	1.14	0.44	1.22	0.68	45

The unexpected observations for version 1 are in Table 33.

Table 33: Unexpected observations (version 1)

Raters	Test-taker responses	Obs	Exp	Res	Var	StRes
R07	V1-119P1	3	1.16	1.84	0.15	4.79
R08	V1-155P3	1	2.86	-1.86	0.13	-5.12
R09	V1-138P3	2	2.94	-0.94	0.06	-3.82
R09	V1-154P3	2	2.94	-0.94	0.06	-3.82
R12	V1-107P3	2	1.03	0.97	0.03	5.7
R12	V1-107P1	1	2.69	-1.69	0.26	-3.33
R14	V1-112P2	1	2.77	-1.77	0.2	-3.94
R18	V1-136P2	2	2.95	-0.95	0.05	-4.1
R19	V1-154P1	2	2.96	-0.96	0.04	-4.71
R27	V1-147P2	2	2.92	-0.92	0.08	-3.32
R27	V1-155P2	2	2.92	-0.92	0.08	-3.32
R29	V1-154P1	2	2.95	-0.95	0.05	-4.15
R29	V1-147P3	2	2.94	-0.94	0.05	-4.06
R30	V1-113P2	2	1.07	0.93	0.07	3.51
R33	V1-155P2	1	2.87	-1.87	0.12	-5.35
R35	V1-125P2	2	2.97	-0.97	0.03	-5.8

Table 34 shows the score and measure table for version 2. Rater 13 is the most severe and rater 06 is the most lenient. Raters 29, 34, 24, 11, 09, 35 and 27 are misfitting but these are not significant.

Table 34: Raters: on-topic score and measure table (version 2)

	A	B	C	D	E	F	G	H	I	J
1	Raters	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
2	R13	1.34	0.33	2.44	2.35	1.23	0.91	0.93	-0.02	45
3	R10	1.13	0.34	2.49	2.44	1.17	0.68	1.15	0.47	45
4	R26	1.13	0.34	2.49	2.44	0.96	-0.08	1.17	0.52	45
5	R33	1.13	0.34	2.49	2.44	0.98	0.02	1.3	0.76	45
6	R29	1.01	0.34	2.51	2.49	1.59	1.91	1.73	1.47	45
7	R34	0.89	0.35	2.53	2.53	1.23	0.84	1.51	1.1	45
8	R23	0.64	0.36	2.58	2.62	0.62	-1.34	0.79	-0.29	45
9	R30	0.64	0.36	2.58	2.62	0.84	-0.47	0.66	-0.61	45
10	R16	0.51	0.37	2.6	2.66	1.19	0.67	1.05	0.26	45
11	R24	0.51	0.37	2.6	2.66	1.51	1.52	1.31	0.71	45
12	R21	0.41	0.4	2.62	2.69	0.89	-0.23	0.56	-0.71	39
13	R11	0.37	0.38	2.62	2.7	1.1	0.41	1.51	0.98	45
14	R25	0.37	0.38	2.62	2.7	0.83	-0.44	1.13	0.41	45
15	R31	0.37	0.38	2.62	2.7	0.57	-1.47	0.45	-1.1	45
16	R17	0.22	0.39	2.64	2.73	0.73	-0.77	0.43	-1.08	45
17	R07	0.06	0.41	2.66	2.77	0.68	-0.91	0.52	-0.75	44
18	R09	0.06	0.41	2.67	2.77	1.73	1.83	2.44	1.87	45
19	R15	-0.11	0.42	2.69	2.8	0.7	-0.79	0.56	-0.55	45
20	R19	-0.29	0.44	2.71	2.83	0.64	-0.95	0.28	-1.17	45
21	R22	-0.49	0.45	2.73	2.86	0.62	-0.99	0.26	-1.08	45
22	R12	-0.7	0.47	2.76	2.89	1.21	0.62	0.93	0.22	45
23	R18	-0.7	0.47	2.76	2.89	0.7	-0.67	0.59	-0.26	45
24	R20	-0.7	0.47	2.76	2.89	0.76	-0.5	0.31	-0.8	45
25	R35	-0.7	0.47	2.76	2.89	0.75	-0.53	1.4	0.7	45
26	R36	-0.7	0.47	2.76	2.89	0.65	-0.86	0.3	-0.83	45
27	R08	-0.92	0.49	2.78	2.91	1.24	0.67	1	0.33	45
28	R28	-0.92	0.49	2.78	2.91	0.6	-0.98	0.21	-0.9	45
29	R14	-1.17	0.5	2.8	2.93	0.56	-1.07	0.28	-0.56	45
30	R27	-1.35	0.54	2.82	2.94	1.89	1.7	1.64	0.85	44
31	R06	-2.02	0.56	2.87	2.97	0.81	-0.29	0.18	-0.2	45

The unexpected observations are in Table 35.

Table 35: Unexpected observations (version 2)

Raters	Test-taker responses	Obs	Exp	Res	Var	StRes	ABS
R08	V2-126P2	2	2.94	-0.94	0.05	-4.04	4.04
R09	V2-125P3	2	2.97	-0.97	0.03	-6.01	6.01
R09	V2-126P3	3	1.3	1.7	0.26	3.35	3.35
R10	V2-115P1	2	2.92	-0.92	0.07	-3.42	3.42
R11	V2-125P1	2	2.96	-0.96	0.04	-5.12	5.12
R18	V2-154P3	2	2.93	-0.93	0.07	-3.66	3.66
R23	V2-101P3	2	1.08	0.92	0.08	3.22	3.22
R24	V2-126P3	3	1.2	1.8	0.18	4.22	4.22
R24	V2-126P1	2	2.92	-0.92	0.08	-3.24	3.24
R25	V2-101P2	2	2.96	-0.96	0.04	-5.12	5.12
R26	V2-145P1	2	1.06	0.94	0.06	3.78	3.78
R27	V2-125P2	2	2.94	-0.94	0.06	-3.87	3.87
R27	V2-154P1	2	2.93	-0.93	0.07	-3.58	3.58
R27	V2-112P2	2	2.93	-0.93	0.07	-3.58	3.58
R29	V2-101P3	2	1.06	0.94	0.06	3.93	3.93
R29	V2-107P1	2	2.93	-0.93	0.07	-3.64	3.64
R33	V2-145P1	2	1.06	0.94	0.06	3.78	3.78
R33	V2-116P2	2	2.92	-0.92	0.08	-3.35	3.35
R34	V2-138P2	2	2.94	-0.94	0.06	-3.88	3.88
R34	V2-129P3	2	2.94	-0.94	0.06	-3.88	3.88
R35	V2-126P1	2	2.97	-0.97	0.03	-6.14	6.14

Table 36 shows the score and measure table for version 3. Rater 13 is the most severe and rater 27 is the most lenient. Raters 29, 33, 31, 14, 11 and 18 are misfitting. Only rater 29 has a significant outfit misfit.

Table 36: Raters on-topic score and measure table (version 3)

	A	B	C	D	E	F	G	H	I	J
1	Raters	Measure	S.E.	Obs.Avge	FairMAvge	InfitMS	InfitZ	OutfitMS	OutfitZ	T.Count
2	R13	2.11	0.32	2.44	2.27	0.9	-0.33	0.85	-0.51	45
3	R34	1.56	0.34	2.56	2.47	1.18	0.76	1.26	0.87	45
4	R07	1.45	0.34	2.58	2.51	0.99	0.06	0.83	-0.42	45
5	R26	1.21	0.35	2.62	2.58	1.14	0.59	1.3	0.87	45
6	R16	1.08	0.36	2.64	2.62	0.92	-0.21	1.13	0.44	45
7	R09	0.95	0.37	2.67	2.66	1.22	0.86	1.01	0.17	45
8	R29	0.95	0.37	2.67	2.66	1.09	0.4	2.04	2.1	45
9	R33	0.95	0.37	2.67	2.66	1.35	1.24	1.4	1	45
10	R25	0.68	0.38	2.71	2.72	0.86	-0.43	0.58	-0.92	45
11	R15	0.37	0.4	2.76	2.78	0.68	-1.08	0.51	-0.94	45
12	R31	0.37	0.4	2.76	2.78	1.04	0.23	1.32	0.72	45
13	R23	0.21	0.41	2.78	2.81	0.59	-1.44	0.58	-0.64	45
14	R14	0.03	0.43	2.8	2.84	1.33	1.02	1.66	1.07	45
15	R24	0.03	0.43	2.8	2.84	0.97	0.03	0.63	-0.45	45
16	R08	-0.17	0.45	2.82	2.87	0.76	-0.64	0.45	-0.75	45
17	R10	-0.17	0.45	2.82	2.87	0.74	-0.69	0.66	-0.31	45
18	R17	-0.17	0.45	2.82	2.87	1.2	0.65	0.73	-0.19	45
19	R20	-0.17	0.45	2.82	2.87	0.66	-0.99	0.49	-0.66	45
20	R30	-0.17	0.45	2.82	2.87	1	0.11	0.53	-0.56	45
21	R36	-0.17	0.45	2.82	2.87	1.23	0.74	0.99	0.22	45
22	R06	-0.38	0.47	2.84	2.89	0.78	-0.52	1.09	0.37	45
23	R11	-0.38	0.47	2.84	2.89	1.63	1.57	0.84	0.05	45
24	R35	-0.38	0.47	2.84	2.89	1.1	0.38	0.82	0.02	45
25	R22	-0.61	0.5	2.87	2.91	0.98	0.07	0.51	-0.38	45
26	R12	-0.88	0.53	2.89	2.93	0.89	-0.13	0.68	-0.02	45
27	R19	-1.19	0.58	2.91	2.95	1.09	0.34	0.49	-0.13	45
28	R21	-1.19	0.58	2.91	2.95	1.2	0.56	0.38	-0.3	45
29	R28	-1.54	0.65	2.93	2.96	0.89	-0.03	0.42	-0.02	44
30	R18	-1.57	0.65	2.93	2.96	1.34	0.77	0.33	-0.16	45
31	R27	-2.86	1.05	2.98	2.99	0.7	-0.06	0.11	0.37	45

The unexpected observations for version 3 are in Table 37.

Table 37: Unexpected observations (version 3)

Raters	Test-taker	Obs	Stp	Exp	Res	Var	StRes
R06	V3-154P2	2	2	2.95	-0.95	0.05	-4.43
R09	V3-136P2	1	1	2.72	-1.72	0.22	-3.65
R11	V3-105P3	1	1	2.67	-1.67	0.26	-3.29
R12	V3-112P3	2	2	2.92	-0.92	0.08	-3.35
R14	V3-146P3	2	2	2.95	-0.95	0.04	-4.55
R14	V3-140P2	2	2	2.9	-0.9	0.09	-3.01
R14	V3-121P3	2	2	2.9	-0.9	0.09	-3.01
R16	V3-113P3	2	2	2.94	-0.94	0.06	-3.87
R24	V3-136P3	1	1	2.66	-1.66	0.26	-3.24
R26	V3-121P2	2	2	2.93	-0.93	0.07	-3.62
R29	V3-105P1	2	2	2.95	-0.95	0.05	-4.14
R29	V3-143P1	2	2	2.95	-0.95	0.05	-4.14
R31	V3-140P1	2	2	2.94	-0.94	0.06	-3.81
R31	V3-154P2	2	2	2.9	-0.9	0.09	-3
R33	V3-121P3	1	1	2.78	-1.78	0.19	-4.12
R34	V3-154P3	2	2	2.9	-0.9	0.09	-3.01
R35	V3-121P3	2	2	2.93	-0.93	0.06	-3.72
R36	V3-136P1	1	1	2.79	-1.79	0.18	-4.26

4.1.6 Bias interaction analysis

I conducted bias interaction analysis to understand if there were any bias interaction patterns between Aptis scores and the on-topic decisions. Bias interaction analyses or ‘differential dimensionality effects’ compare how one facet (or variable) interacts with another facet (Eckes, 2015; Lumley & McNamara, 1995; Myford & Wolfe, 2003). Wigglesworth (1993) describes bias interaction analysis as a way to identify “systematic subpatterns of behaviour which may occur from an interaction of a particular rater with particular respect to some aspect of the rating situation” (p. 309).

There are two types of bias interaction analyses: confirmatory and exploratory. Confirmatory bias interaction analysis has a hypothesis, whereas exploratory bias interaction analysis has no hypothesis (Eckes, 2015, p 132-139). I conducted an exploratory bias interaction analysis as I did not have a hypothesis to confirm. Whereas the previous analyses focused on misfitting raters or test-taker responses,

these exploratory bias analyses combine the misfit findings and identifies bias between particular raters and particular test-taker responses or prompts.

The three-facet FACETS specification file is in Figure 16. The three facets are the raters (facet 1), the test-taker responses (facet 2) and the Aptis scores and on-topic decisions (facet 3). Facets 2 and 3 are positive and facet 1 is negative, as in previous analyses. Facet 3 is non-centred as the focus of the analysis. There are two rating scales and therefore two models. The first model uses the 7-point Aptis rating scale and the second model uses the 3-point on-topic rating scale. There are four columns in the data for facet 3 (Dvalues = 3, 1-4). The first column in the data is the Aptis score, the second to fourth columns are the on-topic decisions. In the models, these are represented by 1B and 2-4B. The bias analysis is between the raters (facet 1) and the Aptis decisions and the on-topic decisions (facet 3), identified with ‘Bs’.

Figure 16: Bias interaction specification file

```
Title = Exploratory bias interaction

Facets = 3
Positive = 2,3
Noncentered = 3
Vertical = 1A,2A,3A
Inter-rater = 1
Arrange = mN
Dvalues = 3, 1-4
*
Models =
?B,?,1B,Aptis Scale
?B,?,2-4B,Ontopic Scale
*
Rating scale = Aptis Scale, R6
0 = 0_<B1
1 = 1_B1.1
2 = 2_B1.2
3 = 3_B2.1
4 = 4_B2.1
5 = 5_C1
6 = 6_C2
*
Rating scale = Ontopic Scale, R3
3 = On-topic
2 = Borderline
1 = Off-topic
*
```

The model has the same rater and test-taker labels. The third facet is the Aptis and the on-topic decisions.

```
3, Decisions
1 = Aptis Score
2 = Prompt 1
3 = Prompt 2
4 = Prompt 3
*
data = Exploratory bias data.xlsx
```

The exploratory bias interaction analysis output provides the *t* statistics (Tables 38 and 39). A bias occurs if the *t* statistic is higher than +2 or lower than -2 (Eckes,

2015; McNamara, 1996). For example, rater 06 has a bias with test-taker response V3-154. Test-taker response V1-112 has a bias with raters 14 and 18.

Table 38: Bias analysis: Raters 06-20

	R06	R07	R08	R09	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20
V1-107	-1.27	1.3	-0.35	0.57	1.33	0.63	-1.89	-1.25	0.65	-0.23	-0.1	-0.64	-0.06	0.38	-1.28
V1-112	0.07	-0.65	1.14	1.91	1.58	1.73	-1.49	0.61	-2.62	0.88	1.89	-1.29	-2.3	0.29	0.63
V1- 113	-0.64	0.63	-0.23	-0.12	0.78	0.56	-0.29	0.07	1.26	0.27	-0.19	-0.75	-1.37	-0.25	-0.22
V1-117	-0.32	-0.78	0.26	-1.64	0.87	0.23	0	-0.26	0.8	-0.35	-1.7	1.03	0.41	0.65	0.11
V1-118	0.75	-0.75	0.63	0.26	-0.22	1.22	-0.18	-1.32	-0.29	0.16	-0.9	-2.46	0.74	-1.1	-0.09
V1-119	-1.31	1.78	0.97	1.1	1.78	-0.58	0.82	0.07	-0.58	-0.85	-0.14	-1.24	-0.1	-1.96	1.5
V1-125	1.72	-0.29	0.52	-0.44	0.3	-2.25	-0.46	0.17	0.21	-0.63	-0.5	0.52	0.68	0.01	-0.32
V1-132	-1.25	-1.67	-0.22	-0.06	0.39	-1.39	0.43	1.26	-0.78	-1.13	0.83	0.56	-0.06	1.34	-0.52
V1-135	-1.1	-0.14	0.6	-0.29	1.22	-0.04	-0.33	-0.78	1.15	0.21	0.33	0.62	-2.01	-0.75	0.57
V1-136	0.98	0.66	0.85	0.58	-1.06	-0.78	0.14	-0.8	0.03	-0.57	0.54	0.27	-0.35	-0.11	-0.9
V1-138	-1.02	-0.67	0.32	-0.83	-0.86	-0.51	-0.8	-0.1	0.86	-0.26	0.61	-0.59	-0.57	0.71	1.05
V1-147	1.01	-1.63	0.88	-2.28	0.15	0.36	0.19	1.01	0.07	-1.53	0.59	0.31	-0.31	1.16	0.27
V1-154	-0.07	0.01	-0.25	-1.88	-0.72	0.58	0.4	1.23	0.29	0.73	0.8	1.59	1.15	-1.01	0.49
V1-155	1.96	0.14	-3.71	1.88	0.46	0.68	1.57	1.33	1.49	-1.02	1.86	-2.27	1.22	0.24	-1.42
V1-157	0.96	0.62	1.58	-0.48	0.07	1.42	0.11	0.03	-1.2	-1.66	-0.53	1.38	0.94	1.84	0.19
V2-101	-0.8	0.2	-1.09	0.05	0.65	0.26	-0.03	0.77	-0.21	-0.14	-0.01	-0.52	0.05	-1.23	-0.6
V2-107	0.58	-0.51	1.5	0.05	1.02	1.23	-0.66	-0.02	0.94	1.39	-0.73	-0.45	1.61	-1.95	1.14
V2-112	1.83	1.25	2.38	-0.15	-1.73	1.51	0.57	-0.62	0.42	0.34	1.09	0.73	-0.09	0.22	-0.79
V2-115	1.32	1.14	-0.03	0.2	-4.33	0.8	0.62	0.71	0.51	0.06	1.02	-1.16	0.13	1.48	0.71
V2-116	-0.28	0.2	-0.58	0.63	-0.24	0.15	0.49	1.36	-0.38	0.44	0.57	0.06	-0.32	-0.65	-0.65
V2-120	0.55	0.05	0.42	-0.03	0.86	-0.29	-0.47	-0.73	0.8	-0.14	-0.07	0.97	0.53	0.7	-0.38
V2-125	1.05	1.76	-0.41	-2.17	0.21	-2.65	0.25	1.07	0.13	-0.44	-1.32	-0.71	-0.25	-1.22	0.33
V2-126	1.26	0.75	-1.42	1.16	-0.28	0.12	-0.21	-0.97	-0.41	0.41	0.53	0.02	-1.13	-0.69	-0.69
V2-129	0.63	0.26	1.53	0.83	-1.46	-1.06	-0.58	-0.49	0.99	-0.04	0.07	1.23	0.61	0.84	1.19
V2-130	1.12	0.86	-0.31	0.77	-2.86	0.52	-0.75	1.17	1.38	0.67	0.74	0.47	-0.15	0.08	0.43
V2-138	-0.71	0.27	0.58	-2.06	-0.28	-0.07	-0.24	0.58	-0.35	0.09	0.16	1.13	-0.74	0.86	-1.41
V2-140	-1.11	0.53	1	-1.65	-0.3	-2.87	1.11	0.35	0.26	0.98	-1.19	0	1.19	-0.71	-0.07
V2-143	-0.67	-2.84	0.61	0.23	1.05	-0.02	-0.2	0.62	0.99	0.13	-0.93	1.16	0.72	0.89	-0.12
V2-145	1.44	1.61	-2.37	0.98	1.08	-0.81	1.15	-1.11	-0.22	0.08	1.38	-1.73	0.23	-1.63	0.76
V2-154	1.61	0.14	0.37	2.25	-0.77	-2.56	-0.7	0	0.91	-0.91	-0.78	1.15	0.53	1.76	0.27
V3-105	-1.41	-1.17	-0.66	0.06	1.03	-1.13	1.06	0.07	0.06	-0.1	-0.71	0.37	-0.45	0.81	1.15
V3-112	0.12	-0.57	-0.06	-1.53	-1.47	-0.19	-0.42	0.13	0.48	0.92	-0.76	0.72	0.1	0.33	-1.27
V3-113	-0.8	-1.04	0.12	-0.41	-1.18	0.95	-0.18	-0.47	-0.33	-0.57	-1.97	0.9	1.41	1.58	0.86
V3-120	0.21	0.38	-1.78	-1	0.07	1.07	0.13	0.95	-0.78	0.04	-1.63	-0.34	1.04	1.28	-0.42
V3-121	-0.59	0.4	0.67	0.32	1.1	1.26	-0.12	0.71	-2.75	0.21	0.28	0.01	0.78	-0.38	-0.03
V3-130	-0.23	0.76	0.93	0.68	0.21	0.42	0.25	-1.44	0.13	-1.43	-1.32	1.48	1.04	-0.01	0.33
V3-135	-1.94	-0.11	-0.38	0.41	-0.36	-0.74	1.27	-0.15	-0.48	0.24	0.35	1.4	-1.98	1.01	-0.16
V3-136	-0.57	-1.32	-1.71	-1.49	0.14	0.48	-0.5	-0.13	-0.71	0.74	0.85	0.4	0.23	-0.98	1
V3-137	1.39	-1.15	1.26	0.32	0.67	2.51	-1.21	-1.88	0.6	-0.67	-0.56	-1.01	0.22	-0.6	-1.07
V3-139	0.73	-0.23	-0.47	0.97	-1.26	-0.86	1.21	-0.28	0.25	0.13	0.92	-0.21	-1.22	0.95	-1.03
V3-140	-0.82	0.17	-0.18	-0.61	0.63	1.6	-0.78	-0.44	-0.23	0.49	-0.68	-0.55	-0.86	-0.47	0.09
V3-143	-1.67	-1.49	0.25	0.57	-0.06	0.23	0	0.44	-1.08	0.43	1.3	1.03	0.41	-0.35	0.11
V3-146	1.08	-1.08	-0.38	1.72	0.24	0.45	0.28	-0.58	-2.11	0.61	-0.3	0.4	-0.22	0.02	0.36
V3-154	-2.84	-0.27	0.88	-0.16	0.15	0.36	0.19	1.01	-1.1	0.52	0.59	0.31	1	-0.07	-0.84

Table 39: Bias analysis: Raters 21-36

	R21	R22	R23	R24	R25	R26	R27	R28	R29	R30	R31	R33	R34	R35	R36
V1-107	0.78	0.79	-0.83	-1.86	1.21	0.34	-0.58	-0.05	0.15	-0.02	0.73	0.63	0.37	0.68	0.68
V1-112	-1.56	-2.49	0.87	0.75	-2.07	0.58	1.29	1.25	0.48	-1.61	-0.57	1.01	-1.42	0.45	0.75
V1- 113	-0.48	0.16	-0.35	0	-0.62	0.29	-1.25	0.03	-1.7	1.04	-0.53	0.53	0.31	1.28	1.17
V1-117	0.86	-1.81	-1.15	-0.57	0.25	0.94	-0.81	0.42	-0.09	-0.18	-0.09	1.38	0.96	0.81	0.27
V1-118	-0.22	-0.22	1.32	-1.17	0.01	-0.15	-0.61	-0.67	1.04	0.26	1.04	1.42	-0.13	-0.27	1.23
V1-119	-1.56	0.16	-0.87	-1.1	0.03	-0.29	-1.23	-0.09	1.25	-0.66	0.7	2.12	0.33	0.05	-0.53
V1-125	1.13	0.3	0.78	-0.87	-0.15	-1.19	0.74	0.68	1.11	-0.44	-0.57	0.97	-0.39	-0.62	0.62
V1-132	-0.69	-0.68	0.75	0.63	-0.38	0.46	0	-0.06	0.36	-0.05	0.36	-0.01	0.48	0.33	0.63
V1-135	1.21	-0.39	-0.5	-0.72	-0.75	0.5	-0.11	0.76	1.19	-0.29	-1.22	-0.23	-0.25	1.16	0.73
V1-136	0.1	-1.06	-0.58	1.46	1.45	0.17	-0.29	0.97	1.27	0.58	-1.1	-0.39	-0.95	0.05	-0.74
V1-138	-0.87	-0.86	-0.27	-0.47	1.16	0.12	-1.5	0.48	1.86	-0.82	-0.01	0.7	1.94	0.87	0.35
V1-147	1.32	0.15	0.51	0.39	-1.77	1.37	-1.53	1	0.12	-2.27	1.3	0.65	-0.89	1.28	-0.69
V1-154	-0.73	0.36	-0.24	0.6	0.59	0.43	-0.03	1.15	-0.76	-1.87	0.33	-0.96	-0.61	-0.8	-0.4
V1-155	-2.58	-0.59	-0.11	0.7	-1.25	-1.51	0.07	1.22	2.22	-1.53	-0.63	-2.38	0.55	1.5	-0.27
V1-157	0.06	-1.11	-0.63	1.44	0.3	0.14	0.98	-1.71	1.25	-2.41	-1.14	-1.44	0.15	0.01	-0.79
V2-101	-0.84	-0.09	1.13	0.97	-1.07	-0.7	1.93	-1.67	1.37	0.05	0.61	0.11	0.76	-0.18	-0.38
V2-107	-1.93	-0.72	-0.88	2.11	-1.12	-1.41	0.49	0.56	0.99	-0.66	0.1	-1.29	-1.39	0.96	0.46
V2-112	-0.27	-0.26	-1.64	-1.88	-1.9	-2.22	-0.91	0.84	-2.42	-0.79	0.48	1.18	0.63	2.1	1.54
V2-115	1.63	0.59	-1.67	-1.03	0.81	0.66	0.19	0.13	0.56	1.06	0.56	-1.41	-1.29	-1.52	-1.03
V2-116	-1.09	-0.24	0.42	0.2	0.18	-0.11	0.55	-1.09	-0.3	0.64	-0.3	0.1	-0.09	-0.35	-0.43
V2-120	-0.51	-0.51	1.11	-0.27	1.02	-0.44	0.57	-0.96	0.84	-0.03	-0.54	-2.38	0.92	-0.57	-0.27
V2-125	1.36	0.21	0.57	0.45	1.52	0.28	-1.45	-0.25	1.34	0.68	0.18	0.71	-0.81	0.15	-0.61
V2-126	1.35	-0.28	0.39	0.78	0.14	0.5	-0.25	-1.12	-0.33	1.17	-0.33	0.66	-0.92	-0.39	-0.46
V2-129	0.2	1.06	-1.48	-1.72	1.29	-1.33	-0.32	0.61	-1.51	0.13	-0.69	0.18	1.15	-1.56	0.52
V2-130	0.3	0.3	1.69	-1.49	0.53	-1.77	-0.09	-0.15	-2.92	-0.17	-0.84	-0.13	1.49	-0.88	-0.48
V2-138	1.01	-0.28	0.08	-0.04	1.18	1.07	-0.68	-0.74	1	1.35	-0.31	1.38	-0.2	-1.65	-1.25
V2-140	1.69	1.06	-0.22	1.39	2.01	1.15	0.5	0.41	1.7	-2.13	-0.35	-1.56	-0.8	-1.1	0.15
V2-143	-0.24	-0.24	1.3	1.22	1.21	-0.17	-0.63	0.72	-0.27	0.24	-0.27	-0.85	-0.15	-1.59	-1.2
V2-145	-1.87	0.51	0.06	-1.3	-0.21	1.21	0.35	-0.39	0.46	-0.54	1.03	0.95	1.23	0.4	-0.18
V2-154	-0.78	0.99	-0.18	2.09	-0.4	-0.65	-2.33	0.53	-2.4	1.46	0.07	-1.97	-0.63	0.93	1.23
V3-105	1.02	-0.71	-0.12	2.12	0.45	-0.59	0.63	-0.45	-0.75	0.07	0.12	0.12	-2.1	0.97	0.47
V3-112	0.55	0.55	0.01	-0.15	-0.17	0.62	0.16	0.1	0.52	1.03	1.59	-0.65	0.64	-0.54	-0.15
V3-113	0.73	0.73	-0.59	0.09	-0.81	-0.14	0.34	0.19	0.71	1.21	0.71	0.44	-0.12	-0.31	-0.79
V3-120	1.49	0.07	0.02	-0.2	-0.22	0.87	-0.57	-0.67	-0.7	0.24	0.02	0.29	-0.49	0.71	-0.2
V3-121	1.1	-0.16	1.36	0.08	1.27	-0.09	0.82	-0.61	-0.19	0.32	-0.19	-3.61	-0.07	-1.48	0.08
V3-130	0.2	-0.93	-0.45	0.45	-2.62	-0.83	1.08	-0.25	0.18	0.68	1.34	0.71	0.29	0.15	-0.61
V3-135	-1.15	-1.14	0.22	0.03	-0.72	0.52	0.84	1.77	-1.19	1.71	0.39	-0.2	1.32	-1.25	0.03
V3-136	1.54	0.84	0.72	-1.39	-0.15	0.25	1.15	-0.61	-1.3	0.91	0.09	0.96	1.63	0.04	-1.39
V3-137	-1.29	1.7	0.17	0.01	-0.91	-0.22	1.41	0.22	0.65	0.32	-0.35	1.18	-0.2	0.62	-0.89
V3-139	-0.48	0.34	0.11	0.66	-0.1	-1.13	0.77	0.71	0.31	0.3	0.31	-0.99	0.45	-0.57	-0.08
V3-140	1.38	0.63	-0.19	0.28	0.26	0.73	0.99	0.03	-0.91	0.66	-0.91	0.08	-0.7	-0.96	0.95
V3-143	0.86	0.87	1.23	-0.58	1.09	0.94	0.47	-1.71	-1.88	-0.19	-1.01	-0.14	-0.83	-0.13	0.26
V3-146	0.23	0.24	0.6	1.56	0.47	0.31	-0.16	-0.22	-0.92	-1.21	0.21	-0.22	-1.85	0.18	0.48
V3-154	0.14	0.15	0.51	0.39	0.37	0.22	-0.25	-0.31	0.12	0.62	-1.04	-0.34	0.23	-1.08	-0.69

I conducted an additional bias investigation on overall Aptis scoring and prompt decisions. The findings are in Table 40. Rater 34 has bias in their Aptis scoring

decisions and for prompts 2 and 3. This is the rater who was severe with all on-topic decisions and then lenient with overall Aptis speaking scores.

Table 40: Bias interaction t-statistics for Aptis score and prompts

	Aptis Score	Prompt 1	Prompt 2	Prompt 3
R06	0.78	-0.25	-0.69	-0.52
R07	-1.83	-0.48	0.88	2.22
R08	1.63	-2.32	-0.14	-0.49
R09	-1.11	0.59	-0.48	1.46
R10	0.78	-0.21	0	-1.17
R11	-0.62	0.77	0.34	-0.36
R12	-1.53	1.28	1.4	-0.66
R13	-1.11	1.76	-1.32	0.88
R14	-0.85	-0.81	1.21	1.22
R15	-0.97	1.48	-0.99	0.84
R16	0.03	0.65	-1.86	1.36
R17	0.02	-0.72	0.39	0.46
R18	1.9	-0.99	-2.2	-0.1
R19	0.14	-0.91	0.73	0.07
R20	-2.16	0.1	1.85	1.35
R21	0.15	-0.71	0.44	0.17
R22	0.19	0.06	-0.88	0.68
R23	-0.47	0.96	-0.8	0.49
R24	-0.47	0.15	1.23	-0.92
R25	0.8	-0.4	-0.67	0.01
R26	0.42	-0.67	0.85	-0.91
R27	0.23	1.39	-1.95	0.05
R28	0.28	-0.33	-0.06	-0.1
R29	2.96	-0.88	0.48	-4.74
R30	-0.56	0.61	0.5	-0.55
R31	2.76	-1.14	-0.41	-3.07
R33	-0.85	-0.14	0.88	0.45
R34	4.63	-2.38	-3.51	-0.5
R35	-1.54	0.35	2.06	-0.25
R36	-2.96	1.36	2.22	0.39

Raters highlighted in Tables 38- 40 are making significantly biased decisions between on-topic decisions and Aptis scoring compared to Rasch model expectations.

4.1.7 Quantitative analysis conclusion

From these quantitative analyses, I identified raters and test-taker responses for further investigation in the mixed methods interpretation based on leniency/severity, fit statistics, unexpected observations and bias statistics (Tables 41 and 42). It was not possible to do bias interaction for separate prompts or separate versions as the bias analysis needed all three prompts to be able to compare the on-topic decisions with the Aptis scores.

Table 41: Quantitative analysis overview of raters

Analysis focus	Score and measure table	Severe	Lenient	Significant misfits (infits and outfits)	Unexpected observations	Bias
Raters: Aptis scores	Table 12	R13	R08	R21, R24	Table 14	Tables 38-40
Raters: On-topic decisions (separate prompts)	Table 15	R13	R28	R09	Table 20	
Raters: On-topic decisions (combined prompts)	Table 22	R13	R28	R34	Table 25	Tables 38-39
Raters: On-topic decisions for prompt 1	Table 26	R29	R14	None	Table 27	Table 40
Raters: On-topic decisions for prompt 2	Table 28	R13	R35	None	Table 29	Table 40
Raters: On-topic decisions	Table 30	R07	R27	R09 and R30	Table 31	Table 40

for prompt 3						
Raters: On-topic decisions for version 1	Table 32	R13	R08 and R35	None	Table 33	
Raters: On-topic decisions for version 2	Table 34	R13	R06	None	Table 35	
Raters: On-topic decisions for version 3	Table 36	R13	R27	R29	Table 37	

Table 42: Quantitative analysis overview of test-taker responses

Analysis focus	**Score and measure table**	**Severe**	**Lenient**	**Significant misfits (infits and outfits)**	**Unexpected observations**	**Bias**
Test-taker responses: Aptis scores	Table 13	V1-107, V3-136	V1-155	V1-155, V1-119, V2-145	Table 14	Tables 38-39
Test-taker responses: On-topic decisions (separate prompts)	Tables 16-19	V1-107 (prompt 3) in Table 19	Table 16	V1-112 (prompt 2), V2-126 (prompt 3), V1-119 (prompt 1)	Table 20	
Test-taker responses: On-topic decisions (combined prompts)	Table 23	V2-145	V2-120, V2-138	V1-119, V2-145	Table 25	Tables 38-39

4.2 Qualitative analysis results

The qualitative analysis examined the on-topic written perceptions from raters using comparative thematic analysis and NVivo and Excel software as explained in the qualitative analysis procedures in section 3.6.2. The analysis took part in five steps: (1) coding raters' perceptions of test-taker responses (two rounds); (2) second coding a sample of 10% of rater perceptions; (3) recoding the raters' perceptions and conducting frequency counts; (4) collapsing (merging) codes and identifying five internal criteria categories that raters use to make their on-topic decisions; (5) and documenting raters' on-topic decision-making severity, consistency and construct relevance. This section presents the results.

4.2.1 Coding rater perceptions of test-taker responses

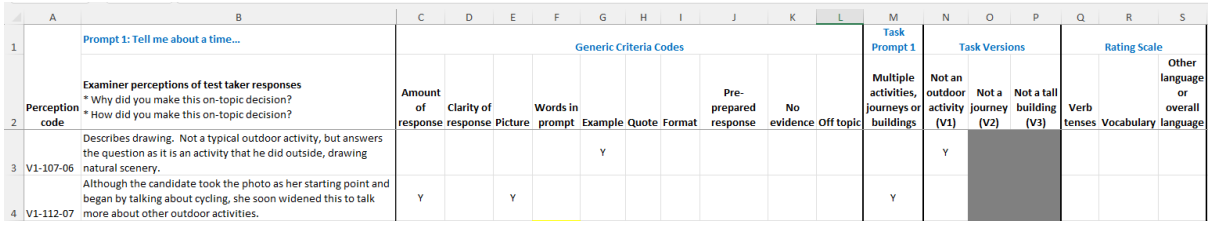
The first analysis focused on coding the raters' perceptions for each test-taker response and was in two rounds, with the second round fine-tuning the codes from the first round (see section 3.6.2.1 for the procedures I used to code the rater perceptions).

Raters ranged in their total length of perception text. Rater 25 had the lowest total length of perception text: 540 words or an average of four words per perception. Rater 10 had the highest total length of perception text: 7654 words or an average of 57 words per perception.

The codes used in round 2 coding are in Figure 17 and Appendix 15. They are explained in section 3.6.2.1.2. The second coder also used these codes.

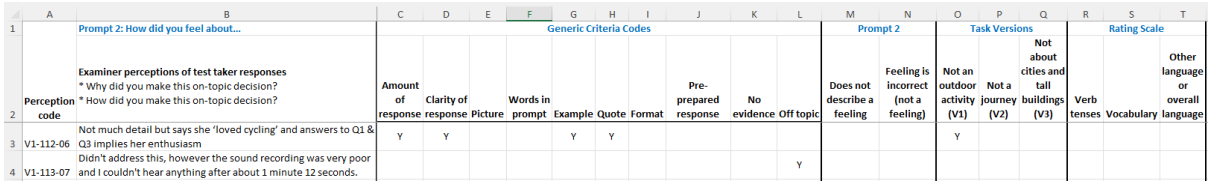
Figure 17: Coding rater perceptions of test-taker responses (round 2)

Prompt 1:



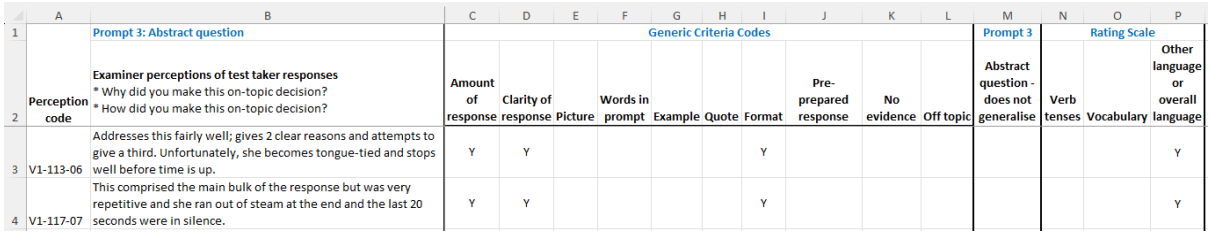
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1		Prompt 1: Tell me about a time...	Generic Criteria Codes										Task Prompt 1	Task Versions			Rating Scale		
2	Perception code	Examiner perceptions of test taker responses * Why did you make this on-topic decision? * How did you make this on-topic decision?	Amount of response	Clarity of response	Picture	Words in prompt	Example	Quote	Format	Pre-prepared response	No evidence	Off topic	Multiple activities, journeys or buildings	Not an outdoor activity (V1)	Not a journey (V2)	Not a tall building (V3)	Verb tenses	Vocabulary	Other language or overall language
3	V1-107-06	Describes drawing. Not a typical outdoor activity, but answers the question as it is an activity that he did outside, drawing natural scenery.					Y							Y					
4	V1-112-07	Although the candidate took the photo as her starting point and began by talking about cycling, she soon widened this to talk more about other outdoor activities.	Y		Y								Y						

Prompt 2:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		Prompt 2: How did you feel about...	Generic Criteria Codes										Prompt 2		Task Versions			Rating Scale		
2	Perception code	Examiner perceptions of test taker responses * Why did you make this on-topic decision? * How did you make this on-topic decision?	Amount of response	Clarity of response	Picture	Words in prompt	Example	Quote	Format	Pre-prepared response	No evidence	Off topic	Does not describe a feeling	Feeling is incorrect (not a feeling)	Not an outdoor activity (V1)	Not a journey (V2)	Not about cities and tall buildings (V3)	Verb tenses	Vocabulary	Other language or overall language
3	V1-112-06	Not much detail but says she 'loved cycling' and answers to Q1 & Q3 implies her enthusiasm	Y	Y			Y	Y							Y					
4	V1-113-07	Didn't address this, however the sound recording was very poor and I couldn't hear anything after about 1 minute 12 seconds.										Y								

Prompt 3



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		Prompt 3: Abstract question	Generic Criteria Codes										Prompt 3	Rating Scale		
2	Perception code	Examiner perceptions of test taker responses * Why did you make this on-topic decision? * How did you make this on-topic decision?	Amount of response	Clarity of response	Picture	Words in prompt	Example	Quote	Format	Pre-prepared response	No evidence	Off topic	Abstract question - does not generalise	Verb tenses	Vocabulary	Other language or overall language
3	V1-113-06	Addresses this fairly well; gives 2 clear reasons and attempts to give a third. Unfortunately, she becomes tongue-tied and stops well before time is up.	Y	Y					Y							Y
4	V1-117-07	This comprised the main bulk of the response but was very repetitive and she ran out of steam at the end and the last 20 seconds were in silence.	Y	Y					Y							Y

I also collected memos, or personal reflections, as I coded (see Figure 18 for an example). I took notes on any patterns or differences of perceptions as I coded and documented quotes to support my perceptions. There were 135 analyses (45 test-taker responses x 9 prompt/version combinations).

Figure 18: Memos while coding

Prompt 1 Version 1: Tell me about an outdoor activity you did as a child.

Test taker response: V1-107

- Wide variation of perceptions between raters who had internal criteria that painting mountains and forests and calligraphy are not outdoor activities.
- Some raters focus on the function of the question about describing (explaining / says) while others focus on 'what is an outdoor activity' and ignore the function.
- The importance of the prompt word 'outdoor':
    - "Describes drawing. Not a typical outdoor activity, but answers the question as it is an activity that he did outside, drawing natural scenery." (R06)

- “The candidate initially talked more about indoor activities but did come round to talking about outdoor activities in terms of doing his painting outdoors.” (R07)
- TT eventually said 'outside'”. (R11)
- “Because TT mentions doing caligraphy outside his house but not about outdoor activity. TT misunderstood outdoor activity as any activity like writing”. (R12)
- “This was a little tricky to begin with. The question clearly asks about "outdoor" activities. While drawing and calligraphy are not "outdoor" activities, the speaker does talk about drawing "outside." I was debating with myself if I should grant the speaker the benefit of the doubt because in a sense he has literally responded to the question. However, on the second listen, I decided that that understanding what an outdoor activity is, is a nuance we should expect speakers to understand. So I chose "off-topic."” (R13)
- “TT describes painting / drawing as their favourite childhood outdoor activity while the picture hints at a more sportive type of activity.” (R17)
- “He talks about drawing the forest/mountains/sea outside.” (R18)
- “Though 'outdoor' implies something more active than drawing, the response successfully explains how the drawing took place in the mountains,forests etc.” (R21)
- “talks about the activity of drawing - outside, in detail”. (R22)
- “Because drawing and calligraphy is normally done indoors”. (R24)
- “met criteria of an activity that was done outdoors” (R25)
- “He spoke about an activity that he did outdoors, rather than an outdoor activity.” (R29)
- “Drawing/painting is not an outdoor activity but can be done outdoors as well.” (R30)
- “He starts this off topic but comes around to talking about being in nature etc so he interprets the question differently to most as he choses a stationary 'activity' outdoors instead of an active one but it's still relevant.” (R35)

Test-taker response: V1-112

- Most raters feel the response addresses the question prompt:
  - “Candidate makes reference to key words in the question and the picture.” (R09)
  - “The speaker gives quite a lot of relevant information.” (R14)
  - “Speaker directly addresses the prompt, describing what she did.” (R16)
  - “Talks about cycling - an activity that can only be done outdoors, and describes how she did it, who with etc.” (R31)
- Some raters did not mention that it was a full response:
  - “TT mentions cycling with the brother” (R12)
  - “speaks about cycling”. (R15)
  - “TT refers to riding a bicycle, racing with brother and friends”. (R17)
- Two raters noted the pictures:
  - “Although the candidate took the photo as her starting point and began by talking about cycling, she soon widened this to talk more about other outdoor activities.” (R07)
  - “Talks about cycling in childhood as in the picture.” (R36)
- Three raters noted that the test-taker spoke about multiple activities.
  - “Although the candidate took the photo as her starting point and began by talking about cycling, she soon widened this to talk more about other outdoor activities.” (R07)
  - “tells about various activities” (R22).

- “TT talks about outdoor activites (sic) she did as a child”. (R23)

- Three raters mentioned clear or clearly:
  - “Clearly describes cycling as a child.” (R06)
  - “Speech is clear and easy to understand” (R09).
  - “Test taker clearly describes what they did”. (R26)

4.2.2 Second coding

The second coding used the codes from round two in the previous section (Figure 17 and Appendix 15 and further explained in section 3.6.2.1.2).

Overall, the exact intercoder agreement meets the 75% minimum threshold suggested by Mackey & Gass (2015) (Table 43). Prompts 1 and 2 also meet the threshold but prompt 3 falls below the suggested minimum for good intercoder reliability.

Table 43: Exact intercoder agreement

	Prompt 1	**Prompt 2**	**Prompt 3**	**Overall**
Total codes	184	174	156	514
Coding agreements	156	131	114	401
Exact intercoder agreement	.85	.75	.73	.78

The findings in Table 43 provide the exact intercoder agreement after some codes were collapsed (merged) and errors rectified. During the analysis of the second coding, it became clear that not all codes worked as expected. There were systematic differences in subjective opinion, which sparked interesting discussions about the best way to interpret the qualitative data. There was also an occasional tendency for both coders to forget they were reviewing the raters’ perceptions and not the test-takers’ responses. There were also ambiguous perceptions that

challenged both coders. Coder drift may have occurred. The coding was rectified to fix a small number of obvious errors and the data in Table 43 reflects these changes. In the next thematic analysis section, I explain the coding results, including why some codes were collapsed (merged).

4.2.3 Thematic analysis

The third part of the analysis provides the results for the coding and begins to answer RQ 1 and sub-research questions 1.1, 1.2, 1.3 and 1.4. See section 2.4 for the research questions, section 3.6.2.3 for the data analysis procedures for the thematic analysis, Appendix 15 for the codes and section 3.6.2.1.2 for a description of the codes.

4.2.3.1 Generic criteria (RQ 1.1)

This section discusses findings for codes that were common across all prompts and versions and addresses research question 1.1: *What generic criteria do raters report using that are common across all speaking test prompts and versions used in this research?*

Amount of response:

Rater perceptions about the expected amount of information provided by the test-taker was generic across all prompts and versions. All raters except one (rater 20) provided evidence that they have internal criteria that believes a response should have a particular depth or detail. I coded any comment by a rater that reported depth or detail including synonyms, antonyms and derivatives (e.g., details, a good amount of time, at length, extended response, fully answered, brief, not enough detail) using the ‘amount of response’ code. For example:

- “This wasn't covered in the depth I was expecting actually.” (R07)
- “Candidate gives plenty of details…”. (R09)

There were 750 instances of words used by raters which referred to ‘amount of response’ as seen in Table 44.

Table 44: Amount of response code frequency count

Amount of response sub-code	Prompt 1	Prompt 2	Prompt 3	Total
detail, details, detailed	120	40	39	199
brief, briefly	38	50	32	120
expand	28	19	35	82
short	17	18	40	75
develop, developed, develops	2	9	55	66
full, fully, fullfilled	15	7	29	51
only	15	19	2	36
length	10	7	11	28
fulfil	1	2	17	20
depth	5	4	8	17
sufficient	7	4	4	15
once	1	7	4	12
argument	0	0	9	9
amount	3	2	1	6
extended	3	1	2	6
small	4	1	0	5
one sentence	1	2	0	3
Total	270	192	288	**750**

All raters except rater 20 referred to the ‘amount of response’. Rater 16 used the ‘amount of response’ the most often to make on-topic decisions (62 times). Raters referred to ‘amount of response’ more frequently for prompts 1 (270 times) and prompt 3 (288 times) than for prompt 2 (192 times). The sub-code ‘detail/details/detailed’ was used the most by raters (199 times), in particular for prompt 1 (120 times).

There were some perceptions that were unclear whether the rater was referring to the ‘amount of response’. For example, the perception “vivid description of trip to Malaysia” (R06) implies that there is sufficient amount of response. The word ‘mentions’ implies that there is not enough response (e.g., “TT mentions cycling with the brother” R12). I did not include these unclear perceptions in the frequency counts. In addition, in a small number of perceptions, more than one of these frequency count words appears, for example “speaks about her experience riding a bike, very brief, does not go into detail” (R19). I decided to keep both references to ‘amount of response’ in the frequency count. These two coding difficulties do not change the finding that ‘amount of response’ is an important criterion raters use for making on-topic decisions and that raters report very different views on the importance of ‘amount of response’.

Clarity of response:

Clarity of response are perceptions that focused on how important it was to raters that the test-taker’s response was clearly spoken and explicitly answered the question prompts. I used the ‘clarity of response’ code for rater perceptions that mentioned ‘clear’ (including antonyms and derivatives), ‘implied’ or ‘explicit’ and ‘fluency’ (including speech speed and pauses).

There were 250 perceptions in the ‘clarity of response’ code (Table 45). The word family with the highest frequencies was ‘clear’, including ‘unclear’, ‘not clear’, clearly, clarity, clarified, clarifies (185).

Table 45: Clarity of response frequency count

Clarify of response sub-code	Prompt 1	Prompt 2	Prompt 3	Total
clear, unclear, not clear, clearly	57	51	77	185
imply/implied	0	6	10	16
explicit	0	4	7	11
understand	5	2	3	10
success	3	0	4	7
hesitation, hesitate, hesitates	0	1	3	4
speed	0	3	0	3
signpost	2	0	1	3
clarity, clarified, clarifies	3	0	0	3
fluency, fluent	0	0	2	2
pause, pausing, pauses	1	0	1	2
intelligible, intelligibility	1	0	1	2
pronunciation	0	1	0	1
flow	0	1	0	1
Total	**72**	**69**	**109**	**250**

Twenty-seven raters referred to ‘clarity of response’ in their perceptions. Three raters (R08, R25 and R29) did mention ‘clarity of response’. Rater 21 had the highest number of perceptions (44), followed by rater 13 who had 30 perceptions. Similar to the ‘amount of response’, there were some instances where raters used more than one sub-code in a perception, for example “Speech is clear and easy to understand” (R09).

Raters used the ‘clear’ family of words to refer to both speech clarity and clarity of content. The speech clarity type of perceptions suggest the rater is referring to the comprehensibility, intelligibility or understandability of the spoken response, for example “volleyball - clear, well structured (sic) answer to tell about her sport” (R19). This first type of perception was also used in perceptions to discuss fluency, including the speed of spoken production and pausing. There was also one reference to pronunciation “pronunciation makes it very difficult to fathom whether the candidate responds to this question or not” (R09) Speech clarity, including

fluency and pronunciation, is assessed in the CEFR decision so this finding that some raters are reporting speech clarity is an important finding and may suggest construct irrelevance when making topic relevance decisions.

The clarity of content perceptions occur when the rater is commenting on the test-taker's adherence to the question prompts: "Gave clear example of an outdoor activity, cricket" (R34). This second type of perception was about the ambiguity (or lack of ambiguity) of the test-taker's response with respect to the question prompt.

It many cases, however, it is difficult to decide if the rater meant speech clarity or clarity of content, for example "She clearly answers the question…" (R13). I decided not to separate this code into the two types of clarity (speech clarity and clarity of content) in the qualitative analysis, but I review this finding in the Discussion chapter.

Picture

There is a picture in this task, which is meant to be a visual aid for the test-taker. However, the type of picture and test-takers' reference to the picture had effects on raters' on-topic perceptions. Raters mentioned the picture (or photo) 54 times in the perceptions.

The photo effect was most prevalent in prompt 1 (*Tell me about an activity/event in the past*). This prompt elicits description functions and some test-takers describe the photo. Some of the comments note that it is common for test-takers to describe the photo, for example "responds to the picture instead of answering about personal past = a common mistake" (R21). There were 43 references to picture or photo in prompt 1. Prompt 2 had seven references and prompt 3 had four references.

The versions also had different perceptions. Most of the perceptions that mentioned photos or pictures were in version 1, which is a picture of a child riding a bicycle.

There were 43 perceptions by raters that the photo could be influencing their choice of outdoor activity as a child, for example:

- “The candidate again uses the picture as a prompt to talk about cycling as his activity.” (R07)
- “Describes activity in the picture”. (R15)
- “Talks about cycling in childhood as in the picture.” (R36)

One rater also expected the test-taker to speak about an activity similar to cycling:

- “TT describes painting / drawing as their favourite childhood outdoor activity while the picture hints at a more sportive type of activity.” (R17)

Version 2 only had two perceptions that referred to the photo. The photo is of a group of people taking a long journey across a desert on camels, which is an unusual activity. There were some test-takers who might have experienced this

mode of transport (see Appendix 11 for list of countries in the sample) but none of the test-takers in the sample mentioned this experience as something they had done.

Rater 10 used quotes to support their perceptions. In two perceptions in version 2, this rater quotes the test-taker referencing the photos: "…like I see, what I see in the picture…they used to travel by using camels and they spent months in Sahara but now I think the old people are using or are travelling by planes" (R10) and "I think the picture there is animals in the desert. I want to say that in my country they are very strange but there are horses and they are people to go riding on" (R10). These quotes show that test-takers, even in version 2, are referring to the photos.

Version 3 had some references to the picture and it was because some test-takers had been to the tall building in the photo (Empire State Building).

- “She talks about how she visited the building in the picture”. (R21)
- “Refers to the picture and then refers to their own experience of same building.” (R26)
- “The first question was not quite on-topic as the TT described the picture and then talked about other building.” (R33)

The findings in this section illustrate that photos can influence test-taker responses. In turn, test-taker responses can influence rater perceptions.

Words in prompt

This code performed very similarly to the codes in the ‘task versions’ and therefore this code was merged with the task version codes and is discussed in section 4.2.3.3.

Test-taker words (examples, quotes and vocabulary)

The ‘example’, ‘quote’ and ‘vocabulary’ codes had disagreement between the researcher and second coder. Whereas I coded ‘vocabulary’ only when the rater mentioned the words ‘vocabulary’ or ‘lexis’ (e.g., “Some awkward use of vocabulary”), the second coder coded ‘vocabulary’ whenever the rater put words from the test-taker response into a perception (e.g., “football”, “seaside resort”). I coded these types of perceptions using the ‘example’ or ‘quote’ codes. I discussed the reason for this coding decision with the second coder and they said that they thought the rater was using vocabulary to make their on-topic decision:

> “Regarding vocabulary - for me, vocabulary is at the heart of whether a response is on or off topic. Even at A1 level, vocabulary can still be produced which is on-topic (eg. 'Tell me about an activity you did as a child?' Bicycle, run, game etc). Grammar/cohesive devices etc are the scaffolding which support the lexis (ah, that's true 'Delta' speak!). So for me, a response which mentions any word that is on-topic but not in the prompt, will be vocabulary. Perhaps that explains why I code vocabulary a lot. I wonder though if that's the Examiner in me dominating, rather than the double-coder [second coder]? When I am marking, I listen for lexis first, then the grammatical scaffolding and coherence. So lexis (for me) is key to being on/off topic and grammatical scaffolding and coherence are what determines the CEFR level.” (email exchange with second coder)

I found this quote from the second coder to be very insightful and a key finding, reinforcing the literature on how vocabulary is closely related to on-topic decisions (Lumley, 2005; North, 2003; Sasayama & Norris, 2022).

I decided to combine the ‘example’, ‘quote’ and ‘vocabulary’ codes into one code. Initially, the ‘example’ code were perceptions which used vocabulary from the test-taker to provide evidence of topic relevance, but not as a direct quote. The ‘quote’ code was a direct quote from the test-taker. These codes were performing the same role of using words from the test-taker to make on-topic decisions. Moreover, it was sometimes difficult to distinguish quotes from examples; perceived quotes were not always verbatim. Some perceptions such as “enjoyed seeing animals and felt glad” or “amazing experience” could be in the first person (quote) or second person (reported speech). The feedback from the second coder included: “I am finding this [example/quotes] the hardest of the codes to differentiate between. Thanks for your additional expanations (sic) and (as last time), seeing where you coded something is also very helpful.” (email exchange). However, even with additional training, distinguishing between ‘example’, ‘quote’ and ‘vocabulary’ codes proved difficult.

I named the new code ‘test-taker words’ as this better reflected how the ‘example’, ‘quote’ and ‘vocabulary’ codes were being used in practice by both coders in the same way. Essentially, the raters’ on-topic decisions were based on the words, or vocabulary, used by the test-taker. The raters either used the words as a quote or as an example.

Format (layout of response)

In the main study, most raters did not comment on the layout of the response. In the pilot studies, raters used the layout of the spoken response to make topic relevant decisions but this finding was not replicated in the main study. I had separated the decisions for the main study and most raters wrote perceptions on each question prompt separately. There were a small number of comments regarding the layout of

the response, for example “answers fully at start of response” (R22). Rater 33 commented the most on the layout, for example, “TT answered all questions, the first two quite briefly. He spent most of his time on question one, because of this at one stage it looked like his response would be ambiguous but then he went into more detail towards the end on the last question”. This finding shows that raters do use the layout of the response to make topic relevant decisions, but this research project did not capture the prevalence.

Pre-prepared response

This thesis noted in the introduction how raters can struggle to rate perceived memorised responses, which can affect scoring (Burton, 2018; Cox, 2017; Fernandez, 2018; Gates, 2018) (see p. 17). In this research, there was one response that could be perceived as a memorised response. The test-taker uses an idiomatic expression: “…all work and no play makes Jack a dull boy.”

Twenty-nine raters did not mention this idiomatic expression or perceived it as a positive: “TT gave an in-depth on topic responses (sic) by giving an expended (sic) response including an idiomatic expression” (R33). One rater interpreted the idiomatic expression as the test-taker speaking about work-life balance: “TT says it’s good to balance play with work…” (R28). However, one rater perceived that the idiom was a memorised response: “uses some interesting phrases, but again the response sounds like it is scripted” (R09).

This finding shows that it is possible for raters to perceive idiomatic or memorised text as positive or negative.

No evidence

One rater did not always provide evidence for ‘how’ and ‘why’ they made their decisions, for example “The speaker responds relevantly” (R14) or “The speaker is fully on topic.” For prompt 1, rater 14 uses this approach for 31 of the 45 decisions. For prompt 2, 35 times and for prompt 3, 13 times.

Rater 25 did not provide much evidence, for example “described outdoor activity” or “describes feelings”. I had coded these as ‘no evidence’ but it does show that the rater was making their decision based on if the rater addressed the prompt.

The second coder sometimes coded these responses with the ‘clarity of response’ codes, rather than the ‘no evidence’ code. I reviewed coding and decided to only code perceptions as ‘no evidence’ if there was no information on how and why raters made their topic relevance decisions. There was some ambiguity with perceptions such as “The speaker gives a relevant and balanced response” as it shows that the rater expects a “balanced” response but it is not clear what the rater means by balanced.

Off-topic

This code was used to collect how raters describe off-topic test-taker responses. Examples include “This question is not addressed at all” (R13) or “Speaker doesn't mention anything he did as a child” (R16). This code also collected perceptions of what raters noted is missing in responses, for example “No mention of children in general” (R21).

4.2.3.2 Task prompt type (RQ 1.2)

The second investigation on rater internal criteria for making on-topic decisions was on how the prompt types affected raters' perceptions and answers RQ 1.2: *How do B2-level Aptis speaking task prompt types affect raters' decision-making?*

Multiple activities (prompt 1)

For prompt 1, the perceptions by some raters focused on test-takers discussing multiple activities, journeys or buildings, rather than one as the prompt requires. The frequency count for words that were coded as multiple activities is in Table 46.

Table 46: Multiple activities frequency count

Multiple activities sub-code	Prompt 1
buildings	40
one	33
activities	27
an activity	12
examples	11
journeys	11
many	5
several	5
various, varied	5
lot of	4
different	4
list	3
multiple	1
Total	**161**

For version 1, some test-taker responses discussed more than one activity and some raters noted multiple activities as reasons for their on-topic decisions. Some raters considered it a requirement to speak about one activity in depth, while others did not. For example, for version 1, raters mentioned if more than one activity was discussed:

- “Doesn’t really go into one activity, mentions several”. (R06)
- “did a lot of activities/riding a bicycle,playground/zoo (sic)”. (R08)
- “The speaker mentions riding bicycle, playing on the playground and visited the zoo as his outdoor activities. While the question asks the speaker to talk about "an" activity, I would accept this list as on-topic.” (R13)

In version 2, raters noted if more than one journey was discussed:

- “Candidate gives some examples of journeys she has made in and out of her own country.” (R09)
- “Briefly mentions a journey with a friend, although talks more about journeys / distances in general”. (R31)

In version 3, some raters noted if multiple buildings were discussed as a reason for on-topic perceptions.

- “talks about multiple buildings, none in particular”. (R06)
- “Doesn't speak about one specific building”. (R31)

It was clear from rater perceptions that some raters took the requirement to speak about one activity, journey or building seriously. However, as with all codes in this research study, the sample influences the findings and must be used with some caution. I purposely chose some test-taker responses that did not have multiple activities, journeys or buildings. The finding that ‘buildings’ had the highest frequency is more a reflection of the sample.

Feelings (prompt 2)

Prompt 2 asks test-takers to describe their feelings about the activity or event they described in prompt 1 (“How did you feel about it?”). The prompt is an abstract topic about feelings and had a range of findings. Most raters provided examples of a feeling using adjectives:

- “Exciting /liked it very much/Afraid”. (R08)
- “TT describes how he had good fun and adventure”. (R12)
- “thrilled”. (R15)

Some raters did not give any evidence of a feeling:

- “He says how he felt”. (R14)
- “Feelings were mentioned in passing”. (R16)
- “Talked about some of his feeling”. (R24)

Other raters did not consider the response a feeling:

- “rather than describe emotions, he tells what he thought at the time – this tiny center of business in the midst of a desert, etc., and people slooked (sic) like ants”. (R06)
- “Didn’t really talk about how he felt as such, more about how he progressed as a cyclist.” (R07)
- “Unfortuatelly (sic), he didn’t include a response to this question. He only mentioned that he loved Cricket.” (R30)

- “She did not directly response (sic) to how she felt but said the ‘It was an amazing trip for me’.” (R33)

- “Doesn’t talk about feelings but describes how it was hard and overcame the obstacle and learnt from it.” (R35)

These quotes from the rater perceptions show that some raters may not be sure what a feeling means. There are two definitions for ‘to feel’: (1) to experience something physical or emotional and; (2) to have a particular opinion about or attitude towards something (*Cambridge Dictionary*, 2023). A small number of perceptions did not consider the second definition as being on-topic.

Some raters used words such as implied or inference for this prompt:

- “‘I loved cycling’. But mainly talks about feelings / benefits in general and personal enjoyment only implied”. (R31)

- “does not explicitly state her feelings, but this could be implied from her answer to prompt 3, that it is something positive.” (R19)

- “We can infer how the TT feels about it (positively), but they don’t directly address this prompt”. (R28)

- “Inference from response to third prompt – tangential”. (R29)

Prompt 2 was the only prompt where raters struggled to provide evidence and used a range of ways to collect evidence. Raters may not fully understand how to measure topic relevance for this type of abstract prompt.

Personal and general

In prompt 3 test-takers are expected to speak about a wider society issue. Coded perceptions included where a test-taker does not generalise and, for example, continues to talk about a personal experience. During the thematic analysis, it became apparent that it was not just prompt 3 where raters mentioned this perceived error about speaking generally or personally. Prompts 1 and 2 require the test-taker to speak about themselves and there were rater perceptions that noted where raters' perceived that test-takers were not talking about themselves. I extended my coding for the new code 'personal and general' to all three prompts.

There were a variety of ways raters referenced the personal and general. The most frequently used words to get their point across were: 'general'; '(im)personal'; 'himself' / 'herself' / 'themselves'; and 'you', 'own'. There were 232 references to 'personal' and 'general'.

Rater 35 had the highest number of personal and general perceptions using the frequently used words (26 times). All raters mentioned at least one of these words at least once.

Below are examples to illustrate rater perceptions regarding the use of personal and general requirements of the prompts to make their topic relevance decisions. They are across all prompts and versions:

Prompt 1: elicits a personal description

- Version 1: "talks about cycling in general not as an activity he did". (R20)
- Version 2: "no personal experience". (R25)

- Version 3: “Discusses personal experience of visiting the Empire State Building.” (R35)

Prompt 2: elicits a personal comment

- Verson 1: “She explained it in a general manner rather than talknig (sic) about her own feelings.” (R24)
- Version 2: “mentions feelings - tired, angry, amazing - but uses impersonal you”. (R31)
- Version 3: “She mentions how insignificant you can feel in amongst tall buildings but more from a general rather than a personal point of view.” (R07)

Prompt 3: elicits a general comment

- Version 1: “Does not address the question at all. Only talks about himself and why he now as an adult, prefers calligraphy”. (R06)
- Version 2: “TT did not directly talk about herself and referred to "you can" but I still had the feeling that she was talking about herself.” (R33)
- Version 3: “"...they need to be proud of themselves…". Talks more about the city in general.” (R36)

The importance of this finding is that it highlights that there are two parts of the rating scale where the personal or general prompt requirement could be scored: in the grammar criterion and the yes/no binary topic relevance decision. It could be that the raters who do not mention the personal or general might be considering a personal or general error under grammar and not as part of the topic relevance rating

requirements. Conversely, raters who mention the personal or general as evidence of being off-topic could be over-penalising test-takers by taking this error into consideration in both the grammar and on-topic rating requirements. This decision could affect scoring.

While prompt 3 in version 1 asks an abstract question that targets the impersonal, (Is it important for children to do outdoor activities?), version 2 is not an abstract question and could elicit lists of transport options (How do people travel long distances in your country?). It could be argued that prompt 3 in version 1 is also concrete as it could be answered with a yes or no response, but raters perceived that test-takers should speak about wider society (I did not include only a yes/no test-taker response). For prompt 2, there were references to “you” in versions 1 and 3, but not in version 2. Again, there were also perceptions where raters were not explicit regarding the personal or general, for example: “This question is not addressed at all” (R13).

4.2.3.3 Task version (RQ 1.3)

There are three versions (or forms) of the task in the current study (see Appendix 2). These are the themes in the three different versions of the task. There were various perceptions around how strictly raters expected the test-taker to adhere to the wording in the prompts. This was the reason I combined the task version and ‘words in prompt’ codes as they had the same findings. Each version has a different theme with prompts that can be broken down into a number of requirements. Raters differed in how important each of these prompt requirements were for speaking on-topic.

Version 1

The prompts for version 1 were:

1. Tell me about an outdoor activity you did as a child.

2. How did you feel about doing this activity?

3. Is it important for children to do outdoor activities?

Raters mentioned the following areas as reasons for a response being on-topic: outdoor and childhood. Raters had different tolerance levels for how closely test-takers covered these two areas.

Some raters did not consider some activities as “official” or outdoor activities:

- Calligraphy: “Because TT mentions doing calligraphy outside his house but not about outdoor activity. TT misunderstood outdoor activity as any activity like writing”. (R12)

- Bike stunts: “The speaker gives a relevant response. The speaker talks about doing bike stunts, which I think is fully allowable as an ‘outdoor activity’ even if unusual – which he acknowledges.” (R14)

- Playing in sand: “Although playing in the sand is not an official activity, I accepted this answer…”. (R19)

- Tai Chi: “He referred to Tai chi as an activity he used to do as a kid. He mentioned this was usually done outdoors, but it doesn’t seem Tai chi is necessarily an outdoor activity.” (R30)

- Table-tennis: “Aslo (sic) mentions bike riding and table tennis late on (not sure if table tennis is outdoor”. (R35)

Raters mentioned childhood as a reason for their topic relevance perception.

- “Doesn’t mention anything about TT’s personal experience during his childhood”. (R12)
- “He doesn’t talk about his childhood.” (R18)
- “Talks about playing football during childhood.” (R36)

Version 2

The prompts for version 2 were:

1. Tell me about a time when you went on a long journey.
2. How did you feel about it?
3. How do people travel long distances in your country?

In this version the key words are ‘long’ and ‘journey’ and raters had different perceptions on the meaning of the word ‘journey’ and on the importance of the length of the journey. A journey is defined as “the act of travelling from one place to another, especially in a vehicle”, “a set of experiences that someone has over a period of time, especially when they change the person in some way” or “a book in which a school child’s achievements and the things they have learned are recorded” (*Cambridge Dictionary*, 2023). All the chosen test-taker responses used the first definition of ‘journey’; it did not occur to me at the start of the study to try and find a sample with the other two definitions. However, the sample included test-taker

responses that did not stick strictly to the dictionary definition of a ‘journey’ and there were differences in the on-topic perceptions by raters to these responses.

Some raters accepted the words ‘trip’ or ‘vacation’ as synonymous with ‘journey’. A trip can also be a journey in which you go somewhere, usually for a short time, and come back again. A ‘holiday’ is a time when someone does not go to work or school but is free to do what they want, such as travel or relax. Whereas a ‘journey’ is focused on the act of moving from one place to another, a ‘trip’ or ‘holiday’ indicate both the movement, the visit and the return. Raters also mentioned the length of the journey as well as the type of transport as reasons for their on-topic perceptions.

There were also some reference to time phrases, for example “She uses time phrases '18 hours' to confirm that she is talking about a long journey” (R09).

Version 3

For version 3, the prompts were:

1. Tell me about a time when you visited a very tall building.
2. How did you feel about it?
3. Why do you think so many cities have tall buildings?

The structure of the Aptis speaking B2 task used in this research (Appendix 2) includes three question prompts that are meant to be increasingly more abstract. Prompt 3 in version 3 was the only prompt that was more abstract. Whereas in version 1 the question prompt 3 can be answered with a yes/no response and in version 2 the question prompt can elicit a descriptive list of transport options, the

prompt in version 3 targets a higher-level cognitive and language ability. The test-taker is required to think of reasons why cities have tall building.

In live Aptis testing, once I was aware of this issue, I tightened up the test specifications in my role as Head of Assessment Development and Delivery at the British Council. We now have quality assurance steps for new speaking tests to avoid eliciting yes/no responses and lists in prompts 1 and 2 and to ensure that prompt 3 is abstract. Wording in speaking question prompts should be standardised as much as possible using clear specifications for item writing.

4.2.3.4 Rating scale (RQ 1.4)

Some raters used other parts of the rating scale when discussing their on-topic perceptions (see rating scale in Appendix 4). In the rating scale, language level should be a separate decision as part of the CEFR decision and not included in the on-topic decisions. The five areas raters are expected to use to make the CEFR decision are grammar, vocabulary, pronunciation, fluency and cohesion. A number of raters mentioned language in their on-topic decision, for example:

- “He only answers this question. Excellent range of complex grammatical structures, cohesion and lexis”. (R18)
- “clear conjunctions used to direct the listener to the response, but vocabulary limitations prevent the candidate from expanding too much”. (R09)
- “Candidate gives a response to this prompt, but due to repetition and some limitations in vocabulary there is some strain on the listener.” (R09)

Grammar – verb tenses

Raters varied on whether to identify verb tense errors as off-topic. The test-taker is expected to speak in the past tense for the first and second prompts and in the present tense for the third prompt. Similar to the personal and general finding, raters who did not penalise for speaking in the incorrect verb tense either did not identify the error or considered the error as a grammar error and part of the CEFR decision instead of an on-topic decision.

There were 53 uses of words ‘verb’ ‘tense’, ‘past’, ‘present’ and ‘future’ and sometimes raters used more than one of these words in a perception:

- “I'd say more off-topic than on-topic due to the fact that she talks about how she feels about it in general in the present tense.” (R18)
- “In my childhood I really love to play cricket - despite tense innacuracy (sic) this is on-topic”. (R21)

Prompt 1 (version 1) has the word ‘child’ in the prompt and raters used this word (and childhood) 300 times. However, it was not always clear if raters were using the word because it was a word in the prompt, or as evidence that the test-taker was speaking in the past tense, for example: “Clearly describes cycling as a child” (R06).

Raters may also have been using the word ‘time’ to refer to the verb tense requirements or because it is a word in the prompt:

- “the time when I went on a long journey was with my family…”. (R11)
- “The last time I visit (sic) a very tall building was in 2012 when I was on a business trip in Chicago... Very clear on-topic intro. Specifies building.” (R21)
- “She talked about a visit to a tall building a long time ago in a city”. (R24)

Prompt 2 asks the test-taker to discuss their feelings about the event or activity which they described in prompt 1 (How did you feel about it?). Raters had different perceptions if the test-taker did not speak about their feelings in the past tense at the time of the event and instead spoke about the outcome of the event or about their feelings in the present tense.

Vocabulary (test-taker words)

As discussed in section 4.2.3.1 (Generic criteria, Test-taker words), the vocabulary code was merged with the 'quote' and 'example' codes into 'test-taker words' code. Most of the perceptions that were coded as 'test-taker words' were regarding the vocabulary the test-taker used, for example: "bus journey" (R20). However, there were 24 instances where raters specified the word 'vocabulary' or 'lexis', for example:

- "A good range of vocabulary and more complex ideas expressed." (R18)
- "Her lack of vocabulary constrained her response as she kept on speaking about the view which the tourists pay for." (R33)

Vocabulary is meant to be assessed as part of the CEFR level:

- B2.2: Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding.
- B2.1: Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding.

However, there is an overlap between the CEFR and topic relevance decisions.

Pronunciation

Pronunciation is mentioned once. The rater has difficulty understanding what the test-taker is saying: “pronunciation makes it very difficult to fathom whether the candidate responds to this question or not”. (R09)

Fluency

Fluency is discussed in the generic criteria, clarity of response (section 4.2.3.1).

Cohesion

Cohesion is measured in the Aptis scale by the assessment of cohesive devices.

- B2.2: A limited number of cohesive devices are used to indicate the links between ideas.
- B2.1: A limited number of cohesive devices are used to indicate the links between ideas.

Coherence is not measured in the CEFR rating scale. There were four comments about coherence.

- “He provides adequate explanation and successfully leads into the next prompt.” (R09)
- “The speaker gives a relevant response and produces a nicely structured response overall, although with pauses, while he tries to relate his past experience to his life in general. As such, he's getting close to going off topic, but manages not to!” (R14)

- "It looks cool" a bit of a lack of coherence in this response.” (R18)

- "TT gave a coherent response based on there being a lack of space.” (R33)

Whereas topic development is expected in the IELTS speaking test in the ‘Fluency and coherence’ and ‘Vocabulary’ criteria, in Aptis the identification of cohesive devices in the language criteria, but there is no explicit criteria for coherence (see the rating scale in Appendix 4). It is not clear if or how raters decide if a response is coherent. Raters may assess coherence as part of the language criteria, as part of the their on-topic decisions or not at all.

4.2.4 Internal criteria raters use to make on-topic decisions

From the coding analysis, broad themes emerged of the internal criteria raters use to make topic relevant decisions. I combined and refined the themes into the main internal criteria. These included: (1) Topic development; (2) question prompt content including grammar expectations, pictures and feelings; (3) vocabulary; (4) linguistic characteristics; and (5) approaches to collecting evidence. These internal criteria that raters use to make topic relevance decisions are expanded on in the Discussion chapter (section 5.1) and briefly described below.

The first internal criteria is ‘topic development’ and includes the findings in the ‘amount of response’ and ‘clarity of response’ coding. There is a wide range of topic development expectations. Some raters have one-word perceptions for the amount of response with no reference to the clarity of the response. Other raters expect all parts of the question prompts to be addressed. Raters had different internal criteria for acceptable amount and clarity of response.

The second internal criteria is the ‘question prompt content’ and includes findings for grammar expectations of the question prompts (verb tenses; singular activities, journeys and buildings; and personal and general task requirements), pictures and raters’ perceptions about feelings in prompt 2.

The third internal criteria is ‘vocabulary’ and includes the examples and quotes raters used in to make their topic relevance decisions.

The fourth internal criteria are the linguistic characteristics raters use to make topic relevant decisions. In particular, clarity of speech including fluency and coherence.

The fifth internal criteria are the approaches to collecting evidence. During the thematic analysis, I identified five approaches that raters used to collect on-topic evidence. Some raters always used the same approach, while other raters used a few different approaches.

1. No specific evidence (e.g., “The speaker gives a relevant response.” R14).

2. Short quotes (one word or short phrases) capturing words that directly answer the main parts of the prompt (e.g., “happy”).

3. More extensive quotes of a sentence or two to put the evidence in context and make it clear the response is on-topic.

4. Referring to the words in the prompt.

5. Implied or inferred meaning of what the test-taker has said by imagining what the test-taker meant to say or interpretation of what test-taker has said by reformulating what the test-taker has said.

4.2.5 Rater analysis

The fifth step was to conduct an analysis of the qualitative data for each rater and document evidence regarding the severity, consistency and construct relevance of their perceptions. See section 3.6.2.5 for the procedures for this analysis and for an example of analysis (Table 7).

4.2.5.1 Rater severity

Rater with lenient perceptions had in common a lack of expectation for the full prompt to be addressed. They focused only on the most important part of the prompt to make their on-topic decisions. These raters looked for key evidence and had a 'can do' approach. They looked for things the test-takers could do, rather than what was missing. These lenient raters gave the "benefit of the doubt" or decided it was "close enough" (e.g., raters 06, 13, 21 and 28). They did not expect detail and did not usually refer to criteria that are covered in other parts of the rating scale (e.g., grammar or fluency).

These raters also used inference and interpretation of what a test-taker meant to say, for example: "Inference more than actually directly answering the question - positive reasons for why children should do outdoor activities seem related to own experiences" (R29). For these raters, one sentence or even one word by the test-taker was enough to be on topic, for example: "Cycling as a child" (R18) and "volleyball" (R22). These raters focused on the test-takers being able to minimally answer the question and did not expect test-takers to address all the words in the prompt

For raters with severe perceptions, there is a pattern of requiring most elements of the prompts to be addressed. They also tended to focus on what was missing.

Severe raters also used other parts of the rating scale to make their on-topic decisions such as grammar, fluency and coherence to guide their decisions. Raters with severe perceptions Severe raters expected a full response with all parts of the prompt addressed.

The prompts have a number of different elements. For example, in prompt 1 for version 1: ‘Tell me about an outdoor activity you did as a child’, a very lenient rater focused only on identifying an ‘activity’ in the test-taker’s response. A severe rater had a wider requirement including up to six elements in this prompt: activity, outdoor, one activity (not multiple), reference to childhood, speaking in the past tense and talking about a personal experience and not in general. Severe raters had higher expectations, were more focused on details and did not just identify the key focus or theme of the prompt. They expected a discussion and topic development, for example: “The speaker mentions riding (as in the image) as his outdoor activity. While this directly answers the question, it is too short for the Task 4 level. The speaker is expected to talk more than just mention the activity” (R13). These raters also mentioned question words such as ‘who, what, where, when or how’ as a way to determine if the topic was developed enough.

Severe raters expected strict compliance with the question prompts. They highlighted picture interference and were sceptical about responses which only described photos. Responses that were inaccurately personal or general were noted. There were some narrow perceptions of feelings. They distinguished between ‘journey’ and ‘trip’ in version 2, and there were some severe perceptions on what constituted an outdoor activity in version 1. Prompts 1 and 2 in version 3 ask test-takers to ‘Tell me about a time when you visited a tall building’ and ‘How did you feel

about it'. One test-taker in the sample had never been in a tall building and therefore could not comment on their own personal experience. Lenient raters accepted this explanation as on-topic but severe raters were less generous.

There were raters in this study who were consistently very lenient in their perceptions (raters 06, 08, 11, 22, 24, 27, 28, 36) or consistently very severe (raters 09, 13, 23, 25, 30, 31, and 35). There are also raters who had no extreme pattern of severity in their perceptions (raters 07, 10, 12, 14, 15, 16, 17, 18, 19, 20, 21, 26, 29, 33 and 34). These raters were either severe or lenient with only a particular prompt or version, or towards one of the marking requirements (for example, an expectation for the test-taker to speak in the correct verb tense). This category also included raters who were neither very severe nor very lenient.

4.2.5.2 Rater consistency

I examined the consistency of the reported perceptions. Apart from the consistency of their severity found in the previous section, I also found two additional types of consistency: 1) consistency in how raters reported their evidence, for example using quotes from the test-taker or referring to words in the prompt; and 2) consistency in the language used, for example frequently using the word 'describes'.

For example, I found rater 25 to be a consistent rater using the same reported evidence style and language. This rater used the words 'describe/describes/described/description', 'talk/talks/talked', 'says', or 'mentions' for 111 of the 135 perceptions and then uses the main theme of the prompt, for example, "talks about outdoor activity/cricket" or "describes feelings". This rater had the shortest average length of perceptions (540 words or an average of four words per perception). NVivo coded rater 25 as quantitative data due to their consistent

and repetitive perceptions (e.g., describes an activity, describes a long journey, describes a building). I coded many of these perceptions as ‘no information’ as the rater has not provided much detail on how or why they made their decisions. However, this rater is making a consistent yes/no decision on if the prompt is addressed. Rater 18 had a similar approach.

Rater 10 was also found to be a consistent rater. This rater used the same approach to collecting evidence for each perception in that they always took long quotes, for example “the most interesting activity I have enjoyed when I was a kid was skiing…I went skiing nearly every year when I was in Elementary school…every winter I went to I went skiing with my family and I even attended a ski camp.” This was the rater who had the longest average length of perception (7654 words or an average of 57 words per perception). Using quotes seems like a good rating strategy because it clearly provides the evidence. However, long quotes are impractical to capture in large scale rating situations unless raters are paid to listen to the response a few times to record a quote.

Rater 17 has a consistently thorough and predictable analysis. This rater repeats words from the question prompt and then provides the key evidence to support their decision, for example “TT describes in detail a long, eventful journey to the coast (about 18 hours by car) with family” (R17).

Rater 28 consistently finds examples and quotes from the test-taker’s response: “rode a bike” or “felt freedom”.

Raters with lower consistency had a variety of reported ways to collect evidence and used a variety of language. They often had different expectations for different prompts or versions. Rater 33 had an inconsistent approach to collecting evidence.

There are some very short perceptions, “TT described her holiday”. There are also long perceptions. There were quotes from the test-takers’ performance. This rater sometimes referred to their own cognitive processes: “I listened closely to see if she was on-topic when she answered the questions.” This rater also referred to all three prompts in their perceptions the most frequently, perhaps finding it hard to separate the on-topic decisions.

Another example of a rater with inconsistent perceptions is rater 34. This rater changes how they report their perceptions from prompt 1 to prompt 2. They consistently refer to the main theme and then focus on the amount of detail for prompt 1, “Spoke about her trip to Malasia. Her description was quite detailed”. For prompt 2, there were more quotes and examples, “Gave examples of how you feel when you play volleyball. Examples: You feel with energy, happy and meet a lot of people.”

Raters with high consistency, making similar decisions across all prompts and versions were R08, R10, R11, R17, R20, R22, R24, R25 and R28. Raters with low consistency with a range of approaches to decision-making and reporting were R12, R14, R18, R21, R23, R26, R33, R34 and R35. The rest of the raters had average consistency.

4.2.5.3 Rater construct relevance

Rater construct relevance is measured in two ways. The first way means that the rater is rating topic relevance and not something else. Rater perceptions that had high construct relevance did not mention other parts of the rating scale or anything that was unrelated to speaking on-topic. Perceptions with high construct relevance looked for examples and quotes to cover the key requirements of the prompt. They

used the prompt words in their perceptions and focused on what the test-taker is able to do using short and clear analysis. The second construct relevance is that they are using the two-step marking system to mark this task: first the rater decides on the CEFR level and then they decide how many responses to the prompts are on-topic to fine-tune this decision. I use the perceptions and email exchanges to decide on the first construct relevance and the marking strategy in email exchanges to decide on the second.

Rater 17 has perceptions that are consistently construct relevant, for example “TT describes playing in the sand as their favourite childhood outdoor activity.” (R17 perception to prompt 1, version 1). “TT mentions being scared (by the road conditions) and happy at the same time (we really had a good fun there, it was really a beautiful place).” (R17 perception to prompt 2, version 2).

Rater perceptions that included other rating criteria (e.g., fluency) in their perceptions were categorised as having low construct relevance. Raters with low construct relevancy also had a narrow understanding of ‘feelings’. Low construct relevance also included raters with unclear decisions (e.g., ‘The speaker gives a relevant response’).

Rater 09 is an example of a rater with low construct relevance in their reported perceptions. This rater uses language (e.g., fluency) in their perceptions and does not provide examples of feelings: “gives some suggestions of emotions and feelings about the activity”.

Rater 25 often does not give reasons for how or why they made the decisions, using the same phrases repetitively with little information about their reasons. The perceptions are very concrete yes/no decisions rather than explaining the reasons.

Although the rating is consistent and shows no evidence of construct irrelevance, there is insufficient evidence to understand the construct relevance.

I perceived raters 08, 11, 17, 22, 27 and 30 to have high construct relevance in their reported perceptions. Raters 09, 12, 14, 23, 24, 26, 33 and 34 have low construct relevance in their reported perceptions. The rest of the raters had average construct relevance.

Marking strategy

I asked each rater in the email exchanges: "*What is your marking strategy? How do you normally go about marking Speaking Task 4?*".

The rating scale for this task used in this research requires raters to make the CEFR level decision before the on-topic decision (see section 1.2.4 for a discussion on this marking construct). According to the email interviews with raters on their marking approach, this does not happen all the time in practice. Raters listen to the performance and need to make these two decisions. In practice raters are arbitrating between the two decisions as they listen to the test-taker's response. Some raters follow the construct and separate the two decisions, and some raters do not.

I discuss how these rating strategies influence on-topic decisions in the mixed methods interpretation. What is noteworthy is that the raters who do not use the two-step approach tend to have mixed profiles (see section 4.3.2). In addition, one rater returned to the project after reflecting on their decisions. Rater 10 transcribed the full test-taker recorded response and then highlighted the evidence.

Perceived emotion when rating

I asked each rater in the email exchanges: "*Was there anything you felt really strongly about? One of the down-sides of email interviews is that it can be difficult to understand emotion so if there is anything where you felt any type of emotion when marking the on-topic criteria [on-topic rating requirements], please let me know. It might help me prioritise decisions as important or not so important.*"

I highlighted key words and anything surprising and reflected on how the emotion might affect on-topic decisions, on-topic perceptions and Aptis scoring. None of the raters had any particular strength of feeling about their on-topic decisions.

4.3 Mixed methods results

The two mixed methods analyses in this section integrate and triangulate the full data set and answer *RQ2: What is the relationship between the internal criteria raters reportedly use to make their on-topic decisions and raters' Aptis scoring and on-topic decisions with respect to rating severity, reliability, consistency and construct relevance?*

The first analysis (4.3.1) focuses on the raters and the second analysis (4.3.2) focuses on the test-taker responses. The purpose of this narrative discussion (Creswell & Clark, 2018) is to understand the relationship between severity, leniency, agreement and disagreement (fit statistics, unexpected decisions and bias) in the objective on-topic decisions and Aptis scores (quantitative data in section 4.1; see Table 41: Quantitative analysis overview of raters and Table 42: Quantitative analysis overview of test-taker responses) with the thematic analysis and severity, reliability and construct relevance of the subjective perceptions (qualitative data in

section 4.2) for a complete interpretation of the data. The findings are further triangulated with information from the email exchanges for a full response to RQ2.

4.3.1 Raters

The first section are the results for raters who were lenient with their on-topic decisions (combined prompts), on-topic perceptions and Aptis scores. The second section are the results for the raters who were severe with on-topic decisions (combined prompts), on-topic perceptions and Aptis scores. I used the logits to classify the quantitative data; positive logits indicated a severe rater. These raters appear to be using their on-topic perceptions to make their on-topic decisions and decide on their Aptis scores. The lenient or severe perceptions and on-topic decisions affect Aptis scores.

The third section discusses raters who do not fit into these two categories and have a mixed profile. Mixed profile raters had lenient Aptis scores but severe on-topic decisions and/or perceptions (and vice versa). These are raters who did not seem to be following the marking construct of using their on-topic perceptions and decisions to make their Aptis score decisions.

The raters are categorised in Table 47. The narrative discussions in this section on these raters also include a comment on their rating consistency and construct relevance, and the data is triangulated using the email exchanges.

Table 47: Rater categories

Raters	Rating category	Aptis scores	On-topic decisions	On-topic perceptions
R06	Lenient	Lenient	Lenient	Lenient

R08	Lenient	Lenient	Lenient	Lenient
R09	Severe	Severe	Severe	Severe
R18	Lenient	Lenient	Lenient	Lenient
R22	Lenient	Lenient	Lenient	Lenient
R27	Lenient	Lenient	Lenient	Lenient
R28	Lenient	Lenient	Lenient	Lenient
R20	Mixed	Severe	Lenient	Lenient
R29	Mixed	Lenient	Severe	Severe
R30	Severe	Severe	Severe	Severe
R31	Mixed	Lenient	Severe	Severe
R34	Mixed	Lenient	Severe	Severe
R35	Mixed	Severe	Lenient	Severe
R36	Mixed	Severe	Lenient	Lenient

4.3.1.1 Lenient raters

Rater 06

Rater 06 has lenient on-topic decisions (-0.48 logits for combined prompts) as well as lenient Aptis scores (-0.93 logits). This rater has negative on-topic logits across all prompts and versions and is the most lenient rater for on-topic decisions for version 2. This rater has significant misfit for Aptis scoring. There is one unexpected observation for combined prompts (V1-155, prompt 3), which leads to a significant bias interaction for this test-taker. Rater 06 has two missing Aptis scores because they scored “4 or 5” (B2.2 or C1), which are high scores and further exemplifies their leniency, and also perhaps a hesitancy to make decisions. This rater is very lenient and appears to be using their lenient on-topic decisions to decide on their Aptis scores.

The bias for test-taker response V3-154 is because this rater marked a 3 (B2.1) for the Aptis score, while the sample of raters marked higher (fair average of 3.85). This rater also marked prompt 2 as borderline on-topic, along with only one other rater (R31). All other on-topic decisions for this test-taker response are on-topic. This borderline on-topic decision for prompt 2 and low Aptis score is uncharacteristic for this lenient rater and therefore creates a bias. The on-topic perception explains their on-topic decision: “interesting, lovely, then again lapses into describing the country (clean, etc) rather than building. It’s not so much that this is off-topic as that she runs out of things to say very quickly so then resorts to talking about her visit to Malaysia in general” (R06), which is a similar perception to rater 31. All other raters mentioned perceptions such as “I felt excited to see this high, huge, vey (sic) tall building” (R11), focusing on the word ‘excited’. It could be that raters 06 and 11 missed this sentence.

On-topic perceptions were also lenient. This rater exemplifies the rater who gives test-takers the “benefit of the doubt” and decides the response is “close enough”:

> “Never really describes his emotions at the time except to say that he remembers it was a beautiful moment. But I’d give him the benefit of the doubt here.” (R06)

> “doesn't give any reasons for building tall buildings but talks instead about why people like to visit (or live in?) tall buildings. Not quite the same thing but close enough perhaps in that it involves the same lexical families”. (R06)

This rater uses words such as ‘clearly’ and ‘detail’, but in a ‘can do’ approach, rather than highlighting what is missing, for example: “Clearly describes cycling as a child.”

These lenient findings were confirmed in the email exchanges:

> “So when I'm marking, I do try to give the candidate the benefit of the doubt on some questions, particularly of the latter type and to realize that if their answer seems to stray a bit far afield it may be only their inexperience with the hypothetical situation rather than any deficiency in their language.”

> “It could be that the candidate doesn't find this particular topic interesting or engaging enough to be able to think of much to say, so in this case, she talks about the next best thing, which is the trip on which you saw the building, the surrounding environment, and so on. If a native speaker answered a question that way -- as many native speakers are wont to do -- I think we wouldn't find it particularly odd. In an exam situation we might think 'oh he's waffling' but in real life speech, well, people do waffle don't they? Now, if she were talking about her pet poodle, or her daughter's ability at making cherry pie, I'd consider it off topic, but as it is, it's close enough to the topic for me to consider it relevant.”

The consistency and construct relevance of the perceptions were adequate, neither high nor low, usually focusing on the prompt words for prompt 1 and then using quotes for prompts 2 and 3. This rater focuses separately on the language and on-topic decisions, as required by the task construct as noted in the email exchanges: “Sometimes my mind wanders and I find myself unsure whether all three responses were on topic (particularly if the candidate is strong, or very weak) so then I listen again, if not to the whole thing, at least long enough to be sure that all responses were on topic” (R06).

In the email exchanges, this rater also noted that they may not be able to describe their rating processes: “I'm not sure I (and I suspect this is true of other people as well) that I know how I go about marking. Decisions get made at lightning speed and I'm not at all sure I can describe my thought processes truthfully and accurately.” This finding corroborates Dörnyei (2007) who wrote that raters may not be able to understand and describe their internal thought processes.

Rater 08

Rater 08 is the most lenient Aptis rater (-1.42 logits) and has negative logits for on-topic decisions (-0.48 for combined prompts). This rater had negative on-topic logits across all prompts and versions. There were no significant Aptis misfits but there were two unexpected Aptis scores (V1-155 and V2-145). There were also two unexpected on-topic decisions (V1-155, prompt 3 and V3-136, prompt 1). The rater had a bias for prompt 1 and test-taker responses V1-155, V2-112 and V2-145. This rater is consistently lenient in their scoring and on-topic decisions and appears to be using their lenient on-topic decisions to decide on their Aptis scores.

This rater has short perceptions for every prompt and version and uses examples and short quotes to cover the key requirements of the prompt, for example: “Cricket” (08) or “Exciting /liked it very much/Afraid” (R08).

The two unexpected Aptis scores (V1-155 and V2-145), which also had bias were because the test-taker did not talk about a personal experience or in general:

> “not addressed -just talked about himself”. (V1-155, prompt 3)
>
> “talked about tourists in general”. (V2-145, prompt 2)

I confirmed adherence to the personal and general requirements of the question prompts in the email exchanges:

> "Quite important as the rubric demands it - YOU did as a child / did YOU feel" (email exchange).

The third test-taker response that had bias was V2-112. The rater marked this response at 6 (C2 level), further exemplifying their rating leniency as all responses in this sample were at B2 level and the fair average for this test-taker response was 3.97. They also marked all the responses on-topic (scores of 3), while the fair average was 2.83. Some raters did not perceive that this test-taker response talked about a journey as exemplified by this perception: "journey not described but long trip" (R25).

For V3-136, the rater decided prompt 1 was borderline but does not clarify why "Old/beautiful/very crowded" (R08). However, reviewing the perceptions from other raters, the test-taker does not name the building so this could perhaps be the reason.

Rater 08 is lenient with Aptis scores, on-topic decisions and on-topic perceptions but does on occasion make severe on-topic decisions if a key part of the prompt is perceived to be missing.

This rater had high consistency and construct relevance in their perceptions. They make the same type of decision every time, mentioning the key information using quotes in short perceptions. There is no reference to clarity of speech (e.g., fluency). Rater 08 reports strictly following the two-step marking construct.

Rater 18

Rater 18 is lenient in their Aptis scoring (-1.25 logits) and on-topic decisions (-0.29 logits for combined prompts). This rater has negative logits for all on-topic decisions for prompts and versions except for prompt 2 (0.06 logits) and version 1 (0.18 logits). There are no significant misfits or unexpected observations for Aptis scores or on-topic decisions but there is bias for two test-taker responses (V1-112 and V1-135). There is also a bias for prompt 2.

This rater reports very lenient perceptions and focused on any evidence that answers prompt 1 and 3, often with only one-word perceptions (e.g., “Cricket”) or short phrases (“Because people want to have nice views.”).

Prompt 2 is more complicated and is the reason for the bias findings for V1-112 and V1-135, prompt 2 overall and the positive logits for prompt 2 (version 1) on-topic decisions. For these on-topic decisions, the rater is severe about test-takers who speak about feelings in general instead of personally, speaking in the wrong tense (present instead of past tense), if they do not include enough detail or if there is a lack of coherence. For example:

> “I'd say more off-topic than on-topic due to the fact that she talks about how she feels about it in general in the present tense.” (V1-112, prompt 2)

> “"I love to play cricket, everyone love to play cricket." Not enough detail to determine how he felt playing cricket as a child.” (V1-135, prompt 2)

These severe findings for prompt 2 in version 1 are not found in versions 2 and 3.

This rater also provides evidence that they do not always use their on-topic decisions to make Aptis decisions because they focus on language ability and give the “benefit of the doubt”, which were raised in the email exchanges:

Question from researcher: “V2-145: You marked this response a 5 but two prompts were off-topic? Why did you decide to give this test taker a 5?”

Response from rater: “Due to the language features she demonstrates throughout the answer though, this makes me give her the benefit of the doubt on this one. In terms of task achievement, I can't really justify her inability to answer the first question, perhaps she feels it's a personal question? Perhaps she gets carried away answering one of the questions? What I doubt is that she lacks the comprehension skills to understand what is being asked of her. For a mark of 4, 3 responses need to be on topic but I'll give her credit for the level of English she produces.”

In the email exchanges this rater also mentions the subjective nature of measuring topic relevance and that they prioritise language ability over topic relevance:

> “Accurately judging task achievement often feels harder than judging linguistic features because there is sometimes more subjectivity in play when we try to interpret what the TT is trying to communicate. Naturally task achievement is bound to linguistic ability and the latter feels more measurable, which is probably why in best fit or unclear scenarios, I tend to feel more confident in assessing a candidate's global ability by leaning more on the language features produced.” (R18)

This rater has moderate consistency and construct relevance in their perceptions using a range of approaches to marking somewhat depending on the prompt and version. Their rating approach also shows some evidence of construct relevance that they separate their CEFR and on-topic decisions but they are not strict with this separation: “I tend to give more weight to linguistic features over task achievement

so if I had a candidate who I felt was B2.2/C1 and they answered 2 tasks very clearly but only touched on the last one (ie.(sic) a 5 second response) perhaps due to lack of time, I feel I would give them the benefit of the doubt." (R18)

Rater 22

Rater 22 has moderately lenient Aptis scores (-0.16 logits) and on-topic decisions (-0.09 logits for combined prompts). All prompts and versions are negative, except for prompt 2 (+0.33 logits) and version 1 (+0.49 logits). This rater had no misfits or unexpected scores but had bias for test-taker response V1-112.

The reason for the bias for V1-112 was similar to rater 18 with unexpected on-topic decisions for prompt 2 in version 1. The perception is that the test-taker did not talk about themselves and implies that the tense is incorrect: "how it made her feel not mentioned, rather she says why it is good twice" (R22).

This rater has short and lenient perceptions and has the same approach to all perceptions focusing on examples and quotes that provide evidence to meet the question prompt requirements. This rater further shows evidence of their leniency in an email exchange and also that they focus more on language than topic relevance. It also shows that this rater may not be using the two-step decision construct for marking speaking tests.

Question from researcher: "V2-107: You mark prompt 3 as off-topic but then give the test taker a 4. Why did you decide to give the test taker a 4?"

Response from rater: "I would have given a 5 based on language but removed 1 as prompt 3 was off topic. It would have been a strong 4 or a weak 5 but as no prompt 3 I needed to penalise it, and compromised with a 4."

The response from the rater about their rating strategy was: “I look at the language first, then consider that against the questions, if they are fully covered, partially or not covered.” This rater appears to be following the rating construct but makes exceptions if the response is at C1 level. This finding identifies a possible problem with the marking scale for very high-level test-takers who speak off-topic.

This rater has high consistency and construct relevance in their perceptions. They report the same approach to collecting evidence for each perception focusing on examples and quotes to answer the prompt.

Rater 27

Rater 27 has lenient Aptis scores (-0.74 logits) and on-topic decisions (-0.56 logits for combined prompts). This rater is the most lenient rater for on-topic decisions for prompt 3 (-2.76 logits) and version 3 (-2.86) and is severe (positive logits) for on-topic decisions for prompt 2 in version 1. This rater has two missing data for on-topic decisions for prompt 3, no misfits, unexpected data but had bias for one test-taker (V2-154).

The rater bias for V2-154 was because the rater did not view a trip as the same as a journey. For prompt 1: “Family trip to Romania” and prompt 2: “Wonderful - talks about stay/tourist spots. Thinks she has mixed journey with holiday.” The rater marked both of these responses borderline and scored 3 (B2.1), which was uncharacteristically low. Other raters were more generous and equated a ‘trip’ with a ‘journey’.

Rater 27 has consistent lenient perceptions, documenting only the key information using examples and quotes, for example: “excited, challenged”.

The two missing data for prompt 3 could have affected the logit value, but they would likely still be considered lenient if these two decisions were added to the data. It does not affect the validity of this lenient finding. However, in an email exchange, this rater demonstrates difficulty with another prompt 3 on-topic decision. A review of their prompt 3 decisions overall seem high compared to the content of their perceptions. This email exchange below also demonstrates how raters can struggle with making decisions about feelings.

Question from researcher: “Response V3-140: For prompt 3, you mention “You feel special”. Can you go into more detail on how this answers the prompt 3 instead of prompt 2 about feelings.

Response from rater: “I think it answers the third prompt, but it is borderline. She feels special because she is closer to the clouds…. Plus I feel the third prompt should elicit a general rather than a personal response. It would have been better if she had said, ‘many people feel special on top of tall buildings because it is not a normal experience/a novelty…’”

This rater has high consistency and moderate construct relevance in the perceptions reporting key information in their perceptions. They use examples and quotes. Only in prompt three does this rater mention the amount of response, i.e., reference to topic development. The email exchange regarding the rating strategy is unclear if they follow the rating construct. Their process is to eliminate ‘impossible’ scores and narrow down to the ‘best-fit’ score.

Rater 28

Rater 28 has lenient Aptis scores (-0.84) and has the most lenient on-topic decisions (-0.62 for combined prompts). All prompts and versions have negative logits. This rater has no misfits, unexpected on-topic decisions, bias but has one missing prompt 2 on-topic decision. This rater is a reliable rater who is consistently lenient.

Rater 28 also has the most consistently lenient on-topic perceptions in the sample. This rater does not include all parts of the prompt nor refer to any words from the prompts. This rater finds key information in the test-taker response with a consistent approach of finding examples and quotes. This rater occasionally uses interpretation to make their decisions. Rater 28 is another rater who gives the benefit of the doubt:

Question from the researcher: “Response #V3-134: What made you decide to give this a level C?”

Response from rater: “I think I gave the benefit of the doubt on this one. The TT doesn't give a huge amount of complex constructions, however, some of the language is quite accurate, there is some nice vocab ('urbanised'), and especially the pronunciation/pacing/intonation is C-level. On listening for a second time, I'd say this is borderline B2/C1 level, so I could easily give this a 4 on another day.”

Rater 28 is another rater whose perceptions have high consistency focusing on key words in the prompts and with a consistent approach to documenting examples and quotes. The reported perceptions are construct relevant in that there are no references to anything other than the words in the prompt and the words used by the test-takers.

This rater is very construct relevant for their marking strategy. They listen twice, first for the CEFR level and a second time for the on-topic decisions: “With speaking task 4s I tend to listen as long as I need to judge CEFR level - usually between half and 3/4 of the response, but can be the whole response, and can be the first 10 seconds. Once I've decided on level, I listen through again for task fulfilment, skipping sections I don't need to listen to again” (R28).

4.3.1.2 Severe raters

Rater 09

Rater 09 has severe Aptis scoring (+0.82 logits) and on-topic decisions (+0.33 logits for combined prompts). There is no misfit for the combined prompts but there is a misfit for the separate prompts. This rater has severe on-topic decisions for all prompts and versions, except for prompt 1 (-0.12). There are three missing Aptis scores, two unexpected on-topic decisions (V1-117, prompt 2 and V2-140, prompt 1) and bias for four test-taker responses (V1-147, V2-125, V2-138 and V2-154).

In the on-topic perceptions this rater is severe and requires detail, clarity and topic development. This rater also uses language to make on-topic decisions (e.g., range of vocabulary, grammar, fluency, pronunciation and coherence), for example “Candidate gives a response to this prompt, but due to repetition and some limitations in vocabulury (sic) there is some strain on the listener” (R09). This rater does not usually give examples of feelings for prompt 2 with comments such as “gives some suggestions of emotions and feelings” (R09).

The email exchanges confirmed that this rater has severe perceptions about topic relevance and about the requirements for B2 language.

Question from researcher: “V1-113: You mention that the prompt is not fully discussed. What do you think is missing in the response? Why do you give this test taker a 2 (high B1) rather than a 3 (low B2)?”

Response from rater: “For this response the test taker provides a very brief response to the first question and does not expand to tell us more details about the activity or any feelings about that time…. Her vocabulary is good in places, but in parts their (sic) is some strain on the listener due to pronunciation and sound issues…. She doesn’t use many conjunctions to link her responses together and her grammar structures are not always carried out correctly. She doesn’t discuss any feelings about the activity and only gives one reason to why children should do outdoor activities.”

Question from the researcher: “V1-032: You marked all three prompts on-topic yet scored a 3. Why not a 4? What are the “suggestions of emotions and feelings” that you mention in your comment for prompt 2?”

Response from examiner: “…he speaks very generally in places using ‘you’ as opposed to ‘they’ to talk about children…. The candidate talks about cycling now with a reference to the past. As a result it could be suggested that the test taker has not paid attention to the correct grammar tense that he should use and therefore I don’t think that he should be awarded a b2.2 (sic) if cannot recognise past questions. Furthermore, the candidate fails to use clear connections In (sic) his responses which make (sic) the response seem as a rambling as oppose (sic) to a clearly structured response….”

For the unexpected on-topic decision for test-taker response V1-117, prompt 2, this rater reported that it was off-topic with the comment: “no attempt at this question”.

The perceptions from other raters were generally on-topic (23 raters). The two key phrases for deciding this response was on-topic is exemplified by rater 06: “Felt good; it was fun”. The four raters who decided it was borderline expected topic development and used words such as ‘too short’, ‘briefly’, ‘not well developed’. The two other raters who reported this response as off-topic both had a very similar comment to rater 09. They did not perceive that the test-taker had responded to the question at all.

There is an unexpected decision for V2-125, prompt 3. Rater 09 decided it was borderline while all other raters marked it on-topic. Rater 09 referred to fluency and a lack of enough content in the response as the reasons for this decision: “Candidate begins the response to this prompt, but as a result of hesistation (sic) doesnt (sic) complete their response” (R09). This response is to the prompt: *How do people travel long distances in your country?* Most of the other raters accepted as on-topic a list of transport options, for example: “Gives some great examples from the past and the present as methods of transport withi (sic) the country.” While the test-taker directly answers the question, this finding exemplifies the problem with this question prompt in that it is not an abstract question and can be answered using lists.

This rater is also the only one who decided V1-138 (prompt 3) is borderline with the comment “Speaks generally about how outdoor sport is good for people, not necessarily children until the end of the response” (R09). This is a very severe perception.

Again, this rater is the only one who decides V1-154 (prompt 3) is borderline on-topic: “uses some interesting phrases, but again the response sounds like it is scripted. Candidate does attempt to expand on the reasons” (R09). All other raters

marked this on-topic, ignoring the idiom or perceiving it as a positive: “TT gave an in-depth on topic responses (sic) by giving an expended (sic) response including an idiomatic expression” (R33).

The construct relevance for their reported perceptions for this rater is low. They use a range of language features in their perceptions (e.g., range of vocabulary, grammar, fluency). They mention signposting and do not give examples of feelings.

The rating approach used by this rater is the opposite of the rating construct. They decide the on-topic decisions first, and then the CEFR level: “My usual marking strategy is listening to the response all the way through and making a note of whether the test taker has answered all the prompts. I then listen again to see to what extent they answered the prompts, paying attention to vocabulary and grammar and whether the speech is fluent or not. I sometimes make notes using a table given to me by my team leader if there is a particularly challenging responses (sic)” R09).

Rater 13

Rater 13 is the most severe rater with Aptis rating (+1.43) and on-topic decisions (+0.84 combined on-topic decisions). They are the most severe rater for on-topic prompt 2 and versions 1, 2 and 3 decisions. This rater has no misfit or bias but does have four unexpected on-topic decisions (V2-120, prompt 2; V1-119, prompt 3, V3-113, prompt 2 and V3-130, prompt 2). This rater has four missing Aptis scores.

This rater expects detail, is very focused on the prompt requirements and regularly mentions ‘clarity’ and ‘clear’. Rater 13 had a narrow understanding of ‘feelings’ for prompt 2 and expects the prompt 3 response to be a “mature” and “explored” response.

For the unexpected decision V2-120 (prompt 2), the rater decides the response was borderline:

> “The speaker called the experience "nice" and "aweome" (sic) on two different ocassions (sic). He also said he felt it was thrilling - these may not explicitly answer the feelings question, I would still consider them valid. Was not sure so marked ambiguous [borderline] instead of on-topic” (R13).

This is a very narrow and severe decision for prompt 2 about feelings. This rater was the only one who decided this response was not on-topic.

For test-taker response V3-113, prompt 2, there was a similar severe perception about feelings but topic development was also reported as a reason for the off-topic decision: “There is just one sentence where the speaker says "it was amazing." Other than this there is no description of feelings. At Task 4 level, I would expect a more insightful answer and hence have marked this off topic.” Rater 13 was the only one to decide this response was off-topic. Twenty-three raters decided this response was on-topic. The six raters who decided it was borderline used words such as ‘brief’, ‘not clear’, ‘no use of keywords’, ‘only mentions’, ‘doesn’t develop’, ‘feelings were mentioned in passing’, ‘short’ and ‘few examples’.

For test-taker response V1-119 (prompt 3), the rater scored the response on-topic, as did all other raters. This was considered an unexpectedly high on-topic decision for this rater based on their pattern of severe on-topic decisions and deciding prompts 1 and 2 were off-topic. FACETS expected a score of 1.24 for prompt 3 for this rater.

The email exchanges confirm this severity and also discusses how rater 09 approaches weighting, linear responses and feelings.

Question from researcher: “V1-117: You mention that the responses to prompts 1 and 2 are too short to be on-topic. What is missing?

Response from rater: “I guess, in both these responses I was looking for more detail. For Q1, the speaker mentions playing football and later talks about interacting with friends and making fun. I feel like details about where she played, when she did it, how often, her role would add more substance to her response. Similarly, with prompt 2 there is a clearly straightforward answer, she tells us she felt good and lists 2 reasons why…. Here too, it would be nice to know in detail what feeling good entails - happy, excited, energetic might be some examples.”

Question 2 from researcher: “How much weighting do you give to each of the three prompts?”

Response from rater: “While marking, I would give each prompt equal weight. I would be ok, if one was slightly shorter or longer. But in this case, prompts 1 and 2 are answered in 19 seconds, whereas prompt 3 get 1 min 20 seconds.”

Question 3 from researcher: “How important is it that the response is answered in a linear manner (each response answered separately and in order), rather than as an integrated response?”

Response from rater: “I don't think it is important that the answers are in a linear manner. An integrated response would be just as good as long as it was responding to all the 3 prompts with substance.”

This rater also mentioned in the email exchanges: “I noticed only during this trial that the second question in each of the tasks is a "feeling” question. This is very interesting. Because it challenges to (sic) speaker to move into an affective domain” (R13). This insight provides further evidence that raters may need support in how to rate feelings.

This rater has four missing Aptis scores. In three cases, they entered two scores and in the fourth they left it blank. In two of the entries with two scores, the rater could not decide on a 2 (B1.1) or a 3 (B2.1). In the third, they could not decide between a 3 (B2.1) and 4 (B2.2). As the missing scores are low, the findings are still valid; this is a severe Aptis rater. It also suggests that this rater might be an avoidant and dependent rater as Baker (2012) identified.

Rater 23

Rater 23 has severe Aptis scoring (+0.52) and on-topic decisions (+0.26). They are consistently severe for all prompts and versions with no misfits, bias or unexpected decisions.

The rater perceptions identify that this rater expects topic development and focuses on what is missing: “TT talks briefly about their feelings about playng (sic) football - the topic was not very well developed so I decided it was borderline” (R23).

The picture also influences this rater’s reported perceptions: “TT refers back to picture prompt and talks of feelings about riding rather than own experience of outdoor activities” (R23).

This rater also perceives ‘journey’ to be different to the experience at the destination: “TT talks about her positive feelings about where she stayed rather then (sic) saying what she felt about the trip itself” (R23).

The email exchanges identify that the rating severity is also because of language, for example:

Question from researcher: “Response #V3-113: You perceive this test taker to be on-topic but give a score of 2. What is the reason for the score of 2?”

Response to test-taker: “Having listened again to this response, I have decided to stick with my original assessment of band 2 (B1.2) as the TT displays control of simple grammatical structures but makes little attempt to use more complex structures. The TT speaks slowly which is a sign to me of the TT struggling to express their thoughts as their vocabulary resource is limited. There is a lack of cohesion within the response with only simple cohesive devices used.”

This rater reports low consistency and construct relevance in their perceptions. They do not always explain their decisions in prompt 2 or have a narrow perception of feelings. They are very focused on prompt 1, focusing on what is missing.

This rater, similar to the previous one, also marks the opposite of the construct. They first decide if the responses are on-topic: “My normal strategy is to firstly check that the response is on topic and then start paying close attention to the complexity of the language being used and how successful the TT is in producing these structures. This helps to position the response as likely to be B1 or B2….”

These are interesting findings because although the consistency and construct relevance is low, they are a reliable rater.

Rater 30

Rater 30 has severe Aptis scoring (+0.66), on-topic decisions (+0.4 for combined prompts). This rater is also severe for all prompts and versions except for version 3. This rater has significant misfit for on-topic decisions for prompt 3. This rater has three unexpected decisions for on-topic decisions (V1-154, prompt 1; V1-157, prompt 2 and V2-140, prompt 1), bias for three test-taker responses (V1-147, V1-157 and V2-140). There are three missing Aptis scores.

Rater 30 has consistently severe on-topic perceptions and is focused on the words in the prompt, for example: “He went on a long journey with his father” and “She clarly (sic) explained why she thinks it is important for children to do outdoor activities.” There is often not a lot of additional information on how and why they specifically made their on-topic decisions.

For test-taker response V1-154, prompt 1, rater 30 wrote: “I believe she mentioned that she played in the sea/with sand?” and decided this was borderline. Twenty-six raters decided this was on-topic. The other three raters reported this response as borderline because they either did not perceive playing in sand as an outdoor activity: “Although playing in the sand is not an official activity, I accepted this answer, which contributed to the overall mark” (R19) or for topic development reasons: “Mentions playing in sand but does not expand further” (R29).

Test taker response V1-157 had an unexpected decision for prompt 2 and bias overall between Aptis scoring and on-topic decisions. In prompt 2, the rater decided this was off-topic. One other rater agreed with the off-topic decision. Eighteen raters perceived this response on-topic, while ten raters reported that it was borderline. The reason for this difference in perceptions is because some raters felt it did not

address the personal requirement in the prompt: “she doesn't speak about her own experience but in more general term (sic)” (R15). For this test-taker response, prompts 1 and 3 had consensus on being on-topic.

The Aptis scores for V1-157 range from 3 (B1.1) to 6 (C2). Rater 30 scored the response a 3. The unexpected off-topic decision for prompt 2 and the very low Aptis score explain the bias finding.

Test-taker response V2-140 is almost the same finding. For the prompt 1 unexpected decision, rater 30 decided the response was borderline because the test-taker did not speak about themselves: “He talked about the size of his country and how this results in long journeys, but he didn't mention anything about a long journey he had gone to.” Three other raters also perceived the response as borderline for the same reason. Twenty-six raters reported on-topic decisions. The Aptis scores range from 2 (B1.1) to 5 (C1). Rater 30 did not enter a score and wrote: “I'd escalate this.”

For test-taker response V1-147, all raters decided prompt 1 was on-topic. For prompt 2, rater 30 decided it was borderline with this perception: “He said he was good at this activity, but didn't mention how he felt about doing it as a child.” Three other raters decided it was borderline as well with comments including ‘briefly’ and ‘some suggestions’. One rater noted: “He loved it. Does not mention feelings” (R27). Rater 30 had narrow perceptions of what is included in a feeling, compared to the raters in the sample who marked this response on-topic. For prompt 3, this rater decided it was on-topic, while four other raters decided it was borderline because of a lack of topic development.

Aptis scores for V1-147 ranged from 3 (B2.1) to 5 (C1), with rater 30 scoring a 3. The bias appears to be because of the low Aptis score but high on-topic decision for prompt 3. In addition, this rater is significantly misfitting for prompt 3 so is making different decisions than the sample of raters.

This rater could not rate three responses and responded to the email exchange with this comment, which demonstrates caution with difficult to mark test-taker responses for test fairness reasons:

> “I escalated this response as I think despite her sounding fluent, the candidate is off topic and her responses are off-topic (Qs 1 &2) and ambiguous (Q3). Generally, in such cases, I don't feel comfortable penalizing the candidate for their lack of understanding of the question as that is more to do with their reading skill than speaking abilities. Had the test been designed using a reading into speaking approach, I would then see it fair to give a lower score for their lack of comprehension. I do, however, understand this is to control for memorised chunks of language. Given that as an examiner, and an assessment specialist in general, I am very sensitive about test fairness, in such occasions, I choose to play safe and escalate the response for a second opinion.”

This rater’s perceptions had high construct relevance, focusing on the prompt. However, this rater focuses on topic relevance before the CEFR decision: “Well, normally I'd first check whether the response is on task and addresses all parts in the instructions. I then focus on its linguistic features (mainly checking what the candidate can do in relation to each sub-criteria) to decide on a mark.”

4.3.1.3 Mixed profile raters

The previous two sections were clear-cut cases where raters were lenient with Aptis scoring, on-topic decisions and reported on-topic perceptions, and vice versa for severity. These raters appear to be using their lenient or severe perceptions to make their on-topic decisions, and the lenient or severe decisions to decide on their Aptis scores.

There were raters in the sample, however, who had a mixed profile and are discussed in this section. These are raters who did not seem to be following the marking construct. They did not appear to be using their perceptions to make their on-topic decisions and/or were not using their on-topic decisions to decide on their Aptis scores. There are two types of raters: those with lenient Aptis scores but severe reported on-topic perceptions and/or decisions; and those with severe Aptis scores but lenient on-topic reported perceptions and/or decisions.

Rater 20

Raters 20 had severe Aptis scores (+0.57) but lenient on-topic decisions (-0.33). This rater had no misfits and only one unexpected on-topic decision (V3-112, prompt 3). This rater did not have bias for any test-taker responses but had bias for Aptis scoring.

Rater 20 was initially categorised with severe on-topic perceptions. This rater did not think that painting nature was an outdoor activity and did not think cities have tall buildings for views, which are extreme perceptions not found in other lenient raters. However, apart from these severe perceptions, this rater was quite lenient and followed the same pattern of the raters with lenient perceptions in finding minimal

evidence for the key parts of the prompt. The perceptions have high consistency using quotes or examples for the test-takers' spoke performance. No construct irrelevant perceptions were identified.

I first thought that it was the language level that was causing the severe Aptis scores. However, after delving a bit more deeply through the email exchanges, the rater decided that they had been too severe with the Aptis scoring.

Question from researcher: "I was interested in your comment below: 'What I found challenging was responses which try to answer all 3 prompts but having a language level of B.1 (mark 2) Since (sic) 3 questions were answered I took it to B.2.1 in most cases.' (e.g. responses V1-132, V1-136, V2-138, V2-140, V2-154, V3-134, V3-136). What made you think that these responses are at B1 level? Why did you decided (sic) to award a B2 level?"

Response from rater: "I listened to all the responses again and I found that they are B2.2. I am sorry. I don't know why I thought they are B1 level. All the specified test items are at B2.2. There are enough complex structures used accurately. Cohesive devices are used sufficiently to keep it at B2.2. Vocabulary is enough for B2.2 though there was some inappropriate use in some responses. Some pausing is there but does not cause any strain or affect the fluency. All the questions are answered. I hope the change in my marking does not cause any problems for your work. I do have a problem when I mark Speaking and Writing items. Even when the language is of a particular level, if all the questions are not answered I used to bring it down. Recently I was told I need not put so much weightage on the questions being answered. If it is vaguely mentioned it should be accepted as all being on topic. My thoughts

on this is different. We are using language for communication and in the process if the content is not fully communicating the required information how does having a certain level of proficiency help?"

This rater could be categorised as a rater who is lenient with Aptis scoring, on-topic decisions and on-topic perceptions. This email exchange also notes that senior raters might be instructing some raters not to put so much emphasis on speaking on-topic.

This rater has high consistency in their perceptions using mostly examples and quotes for evidence. They use the two-step rating construct, deciding on the CEFR level, and then deciding on the on-topic decisions.

Rater 29

Rater 29 has lenient Aptis scores (-0.89 logits) and severe on-topic decisions (+0.26 logits for combined prompts). This rater is the most severe rater for prompt 1 but is lenient for prompt 2. There is no misfit for combined prompts but there is misfit for separated prompts and four unexpected on-topic (combined) decisions (V2-130, prompt 1; V1-154, prompt 1; V3-143, prompt 1; V2-143, prompt 2; and V2-154, prompt 1). There is a bias for test-taker responses V1-155, V2-112, V2-130 and V2-154. This rater also has bias for Aptis scores and prompt 3.

This rater scored some responses at B2.2 and even C1 level with off-topic or borderline decisions (e.g., V1-119). All three questions should be on-topic to achieve B2.2 so this rater appears to be disregarding their off-topic decisions in some instances when making their Aptis scoring decisions. For V1-119, the rater decided

one prompt was off-topic and another one was borderline. The email exchange clarified that this rater focuses on language.

Question from researcher: “V1-119: Why did you decide on a score of 5?”

Response from rater: Mainly beacuse (sic) he sounds so natural, with false starts which cannot be described as hesitations, for example, and utterances which are not complete sentences which are characteristic of fluent speakers. He also uses a wide range of vocabulary and connectors. I have trouble with speakers of Indian English because some sentences that are grammatically acceptable in Indian English are not completely acceptable in British English, so it can make it difficult to mark. His repetitions and some errors mark him down from a 6, but I felt he was better than a B2. I think the fluency of a speaker tends to make me give a higher score at upper levels than any other factor.”

One reason for the overall on-topic severity and unexpected decisions for prompt 1, version 2 is because this rater does not accept trip/holiday as a journey: “Solely about the holiday/trip - no mention of what happened on the way to the holiday/trip” (R29).

Another reason for the severity is that this rater expects topic development:

Question from researcher: “V1-113: You mentioned that the test taker did not expand on prompt 1. You write abouut (sic) ‘expand’ quite a bit so it seems important to your decision making. How many sentences do you expect from each prompt for it to be expanded enough? Do you have a system of deciding?”

Response from rater: “I generally mean that they have gone beyond the direct answer to the question. So, when asked about how something felt, it is a response that mentions a feeling and why they felt that way and what else was important to them about that experience. Or if it is talking about doing an outdoor activity when they were a child, then a quick 'name only' mention of the activity is not an extended response. I don't really count sentences, but I will go back and listen again to responses where there doesn't seem to have been a balance between answers in terms of length and weight of response.”

Although prompts 1 and 3 are severe, prompt 2 is lenient. The rater does not mention the type of feeling and uses inference to make decisions:

Question from researcher: “Response #V1-119: You mentioned in prompt 2 that you inferred the feeling. Could you write a bit more about this idea of inference? What was the exact inference in this response?”

Response from rater: “I think I mean a cross between inferred and indirect. So, although the question is not answered in terms of 'I felt....', the things that he mentions doing and the reasons why children should go out and do outdoor activities, seem to be recollections of his own times playing cricket and basketball outside, for example. Or that he didn't spend his time tied to the computer as he believes children tend to do these days.”

The email exchange about the rating strategy further exemplifies why this rater is lenient for prompt 2: “I think that question 2 is more difficult to give an extended answer as candidates are asked to describe their feelings which can often be little more than 'amazing'….” Rater 29 reports that they use the two-step rating strategy.

Rater 31

Rater 31 is lenient with Aptis scores (-0.84) and severe with on-topic decisions (combined) (+0.24). There are no misfits or bias and only two unexpected on-topic decisions (V3-154, prompt 2 and V2-130, prompt 1). This rater has bias for Aptis scores and prompt 3.

The perceptions for rater 31 are focused on detail and reference the prompt. This rater expects description and detail (who, what, where, when, how), mentions response layout and notes any overlap between responses to each prompt and highlights personal/general errors. This rater uses quotes and inference as evidence. There is low consistency in the decision-making with a variety of approaches to finding evidence.

This rater ignores their on-topic decisions when making Aptis decisions. They mark primarily on language. They scored five responses at B2.2 or C1 level with off-topic perceptions and scores. A further 13 responses were scored at B2.2 or C1 level with borderline on-topic decisions. The emails confirm that language is important to this rater:

Question from researcher: “V1-119: You marked this a 5 [B2.2] with two off-topic decisions. What part of the language made you decide it was a 5?” [rater wrote on spreadsheet: “based on language level only”]

Response from rater: “Vocab – e.g. develop stamina, physically fit, strong physique, keep yourself fit. Range of discourse markers and cohesive devices – I guess, so I believe that, what happens is, also, if I look at, not only... but... Fairly good range and accurate use of structures – e.g. relative clauses, have a

tendency to. But not completely smooth – a few unfinished thoughts (my biggest reservation about awarding C1 performance), and some awkward usage e.g. gives an exertion."

This rater raises a dilemma of high level test-takers who speak off-topic:

Question from researcher: "V1-135: You mention that you couldn't decide between a score of 3 and 5. What is the reason for not considering a 4?"

Response from rater: "I think I thought it was borderline B2 / C1 in terms of language level. If B2, then the fact that two responses were not very well developed would warrant reducing the mark to 3/5. However, if a candidate is clearly C1 on this task, then there is no option to reduce the marks for off-topic responses, so 5 is "best fit" even if part of the response is off topic (either I have been told that this is the right procedure or perhaps I have made this up!). On second thoughts, however, in real life I probably would have awarded 4/6 given that I thought the language was borderline and also felt two responses were borderline off-topic – so 4 seems a better fit than either 3 or 5 in this case, even though it goes against the usual (/ my usual!) procedure."

The email exchange about the rating strategy further highlights that this rater is focused primarily on language and does not consistently separate the language and on-topic decisions:

"I don't know if I have a particular strategy - I think I just tend to listen and sort of tick off things in my mind as I hear them, or notice things as I go in terms of deciding language level. I often listen more than once to task 4s as they are probably the ones I find hardest to mark. I think I have realised from doing this

project that I am not entirely consistent in my approach, but weigh up my impressions of both language level and relevance together to find the best fit mark. e.g. if the answer to one prompt was ambiguous but the language was clearly B2, I would probably give the benefit of the doubt and award 4/6 (in fact I may be less likely even to spot the ambiguity in these cases), but if I felt there were some aspects of the B2 criteria that were not quite met, as well as there being an ambiguous answer, I would probably award 3/6 (perhaps I would almost even be looking for the ambiguity to help justify my lower mark)."

Rater 34

This rater is a lenient Aptis rater (-1.18 logits) and has severe on-topic decisions (+0.55 for combined decisions). This rater misfits (outfit) for their on-topic decisions, has bias for one test-taker response (V3-105) and there is one missing Aptis score. There are four unexpected on-topic decisions (V2-138, prompt 2; V3-121, prompt 1, V3-154, prompt 3; and V2-143, prompt 2). This rater has bias for Aptis scores and prompts 1 and 2.

The profile for this rater is similar to rater 29. There are eight B2.2/C1 level Aptis scores with off-topic decisions and a further 17 that are borderline. This rater appears to also ignore their off-topic/borderline decisions.

Rater 34 has different expectations depending on the prompt and version. This rater focuses on the prompt, comments on detail and clarity and notes what is missing in test-taker responses. This rater has different approaches to collecting evidence, using quotes for prompt 2 but no words from the test-taker in the other prompts and

refers to language in prompt 3. Finding evidence is not done consistently. This rater has very severe perceptions on prompt 3 about the expected amount of response.

Question from researcher: “Responses V1-112, V1-113, V1-119, V1-135: You gave these responses scores of 4 with at least one response off-topic. Why did you decided to award a high B2 level?”

Answer from rater: “V1-112 - She did not answer the second question however, I think the language produced was higher than a B2. I was not too sure as it would not be fair to give the candidate a low B2 even though she did not answer the second question…. V1-119 - This response was a difficult one to assess as the candidate used a range of complex constructions and used cohesive devices to link his ideas. He focussed on the 3rd question and did not address the other questions. I think it would be unfair to give him a low B2 as his level of English is a C1 and as there is no C1.1 or C1.2, I thought he desreved (sic) at least a B2.2 to compesate (sic) for his high level of English.

For the other two test-taker responses, the rater would change their score to B2.1 but based on a reappraisal of the language.

This rater follows the two-step rating construct deciding on the CEFR level and then makes the on-topic decisions.

Rater 35

This rater had marginally severe Aptis scores (+0.17 logits) but very lenient on-topic decisions (-0.48 logits). There are three unexpected on-topic decisions (V3-121, prompt 3; V3-140, prompt 3; and V2-129, prompt 1) and bias for V2-112. This rater also has bias for prompt 2.

Although the on-topic decisions are lenient, the on-topic perceptions are severe. Rater 35 did not accept unusual outdoor activities, mentions brief test-taker responses, notes tense errors and highlights a lack of detail. This rater does not appear to be using their perceptions to make their topic relevance decisions. However, the rater does make some comments such as: “Doesn’t go into great detail…. But she does give a specific activity” and “It’s brief but the sentence is there.” Therefore, although the perceptions are severe, there is evidence that the rater ignores their perceptions and is comfortable with rating a response on-topic if the basic requirements of the task are met.

The perceptions are not consistent and there is evidence of construct irrelevance for on-topic decisions for prompt 2: “Doesn't talk about feelings but describes how it was hard and he overcame the obstacle and learnt from it.”

The email exchanges identify that the rater gave low Aptis scores even though all three responses were on-topic for language reasons. The rater did not think test-taker responses V2-130, V2-138, V2-140 or V2-143 were at B2.2 level and scored B2.1.

In the rating strategy email exchanges, the rater confirms that they do not use the two-step marking construct (decide on CEFR level and then decide on the on-topic decisions) and instead marks holistically between all criteria.

> “I usually decide quite early on what grade I think I might give it. Then as I listen I adjust it to compensate. If the candidate starts to struggle at some point it goes down and if they take off and suddenly improve after a shaky start I will often ignore the start as it could be nerves. I think I’m quite lenient about how fully they respond to prompts as long as they stay on topic or don’t

> ignore a question completely. If they are C1 level but ignore a prompt completely I may drop a mark. And the same rule applies to most levels. If a candidate is neither fully B2 nor B1 but somewhere in between I often tend to give a 3… sometimes I just can't choose."

Rater 36

This rater has severe Aptis scoring (+0.9 logits) and lenient on-topic decisions (-0.29 logits). This rater has severe on-topic decisions for prompt 1 (+0.18 logits), no bias and one unexpected on-topic decision (V3-136, prompt 1). This rater also has bias for Aptis scoring and prompt 2.

Rater 36 finds the evidence and doesn't expect detail and has consistently lenient perceptions. This rater is focused on prompt 1 and uses examples and quotes. In prompt 3, this rater notes personal/general errors, amount of response and uses inference.

Reviewing the email exchanges with rater 36 on why so many of the on-topic decisions were lenient but the Aptis scoring decisions were severe, the rater responded by saying they marked the test-taker down because they felt the language was not sufficient to warrant a B2.2 (level 4):

> "V3-112 The candidate fails to respond fluently and adequately; there are noticeable hesitations and pauses while responding to the 2nd & 3rd prompt. Though the level is B2, the candidate does not sustain the same throughout the test" (R36).

Two of the Aptis scoring decisions were considered severe by the rater in retrospect:

> “V3-105 This candidate is definitely at the B2 language level as her vocabulary and sentence structures are quite sufficient for the task and fairly accurate. But, in part 3, she seems to be struggling to respond to the prompt as she talks more about why she prefers cities with tall buildings. I would like to revise her score to 4” (R36).
>
> “V3-154 The response displays good control of vocabulary, structures and cohesive devices; errors are minimal and non-impeding. This candidate's score can be revised to a 4. She fails to get a 5 due to the noticeable hesitations and some L1 influence in pronunciation” (R36).

This rater follows the two-step rating construct.

4.3.2 Test-taker responses

The next mixed methods integration compared the Aptis scoring, on-topic decisions and my thematic analyses of raters’ topic relevance perceptions for each test-taker response. I investigated test-taker responses with the highest (lenient) and lowest (severe) Aptis scores, the most lenient and severe on-topic decisions (I used the combined scores). I used the rater perceptions and email exchanges to triangulate and understand these findings.

The test-taker responses with lenient Aptis scores should have lenient on-topic consensus. The test-taker responses with severe Aptis scores should have severe off-topic consensus. Misfitting Aptis scores and on-topic decisions identify test-taker responses that were not scored in the same way by all raters.

V1-107: Severe Aptis scores

This response was jointly the lowest Aptis scored response, along with V3-136. Three raters marked the response a 0 (below B1), two raters scored the response a 1 (B1.1), eight raters scored the response a 2 (B1.2) and sixteen raters marked it a 3 (B2.1). One rater (R10) scored the response a 4 (B2.2.). Although the majority (17 raters) scored the response at B2 level, the very low scores from some raters reduce the fair average to below B2 level.

Eighteen raters marked prompt 1 on-topic, but the rest of the raters marked it either borderline (9 raters) or off-topic (3 raters). The reason given for marking off-topic by the three raters was the feeling that drawing is not an outdoor activity. Rater 13 explains this decision:

> “This was a little tricky to begin with. The question clearly asks about "outdoor" activities. While drawing and calligraphy are not "outdoor" activities, the speaker does talk about drawing "outside." I was debating with myself if I should grant the speaker the benefit of the doubt because in a sense he has literally responded to the question. However, on the second listen, I decided that that (sic) understanding what an outdoor activity is, is a nuance we should expect speakers to understand. So I chose "off-topic."” (R13)

This perception was queried in the email exchange and the rater responded by saying:

> “As I have mentioned in my perceptions sheet, this was a tricky choice to make. And on reflection I do feel confused all over again and quite inclined to change my response.

> As I understand it, an outdoor activity has an element of physical engagement, for instance, cycling, playing football or running. It is not merely an indoor activity that is done "outside." The image of the child riding a bicycle also reinforced my approach to "outdoor" activities. Hence, I was inclined to not consider this speaker's activity as as (sic) outdoor one.
>
> However, I also do think, as outdoor activity is one that is dependent for its success on to (sic) the outdoors environment it takes place in, for instance, fishing or trekking. The speaker does speak in detail about drawing natural scenery - an act that is dependent on the outdoors environment for its success.
>
> I guess, if my approach to marking is what I say it is in Q 1 - seeing what the candidate can do/ is doing rather than looking for what they can't do, I should award this candidate a mark for this prompt." (R13)

The nine raters who chose 'borderline' acknowledge that drawing may not be an outdoor activity but did not choose off-topic.

By contrast, the eighteen raters who marked the response on-topic were either satisfied with the word 'outside' and noted that the drawing took place as a child, for example: "fulfilled talked (sic) about activity he did as a 10year old outside -drawing" (R08), or they acknowledge that painting is an unusual outdoor activity but still on-topic: "Describes drawing. Not a typical outdoor activity, but answers the question as it is an activity that he did outside, drawing natural scenery." (R06)

Prompt 2 also had a mix of on-topic (18 raters), borderline (8 raters) and off-topic (4 raters) decisions. The main reason raters used for marking borderline or off-topic is expressed by rater 12:

> “Referring to his experience as 'a beautiful moment' doesn't not (sic) clearly describe how he felt about the experience.” (R12)

These raters had perceptions that the test-taker was addressing a memory rather than explicitly answering the prompt about feelings at the time of the activity.

In the email exchange with rater 16, the rater wrote: “Taken literally, I wouldn’t say that necessarily expresses a feeling but I would agree that it could be interpreted as such.” Rater 16 marked this response at B1.2 level because of the two off-topic responses (prompts 2 and 3) but also this rater said in the email exchange that they felt that B2 language was not sustained.

Rater 13 also marked this response off topic because they expected more topic development: “Experiencing a beautiful moment can surely be called a feeling. I did not consider this answer on topic because I found it too general. It was a "beautiful moment" does indicate/ give us a sense of his feelings. But he doesn't really spell out what he was feeling. It is just a one statement answer to the prompt….”

By contrast the 18 raters who marked on-topic wrote something similar to rater 11: “TT says it was a beautiful moment” (R11)

Rater 10, who scored the response on-topic, writes in more detail using a quote:

> “I can remember there are 2 mountains and er very green forest and er sun so I remember that I tried drawing and with different colours and I've kept it these

> papers in my house er I remember it's beautiful moment that I spent then drawing this beautiful scene, ...". (R10)

Prompt 3 is almost consistently rated off-topic (29 raters). Rater 30 gives an explanation:

> "The candidate only mentioned how doing arts as a child resulted in them (sic) growing interest in calligraphy but didn't mention anything about the importance of outdoor activities for children." (R30)

The rater who marked borderline for prompt 3 wrote: "Doesn't answer the 3rd question" (R12), which implies that they also found the response to be off topic. This rater marked the response zero for the Aptis score (below B1 level).

Rater 10 scored prompt 3 off-topic with the perception: "does not mention this at all". However, this rater was the only one who scored B2.2, which should only be awarded if all three questions are on-topic. Reviewing R10's email exchanges, it appears that this rater marks primarily on language and seems to occasionally ignore their on-topic decisions.

The reasons for the low Aptis scores for this response is likely due to topic relevance decisions and possibly language. Some raters do not consider drawing nature scenes as an outdoor activity for prompt 1 and do not consider the perception 'a beautiful moment' to be a feeling for prompt 2. However, there was consensus that prompt 3 was off topic.

V3-136: Severe Aptis scores

Raters gave low Aptis scores and it appears to be in part because of the language level. There were three scores of 1 (B1.1), 17 scores of 2 (B1.2), seven scores of 3 (B2.1) and 3 scores of 4 (B2.2).

For prompt 1, 22 raters marked it on-topic, seven raters marked it borderline, and one rater marked it off-topic. The main reasons for lack of consensus with on-topic decisions were the perceptions that there was a lack of topic development and that there were vocabulary limitations.

Examples of comments for raters who marked this response on-topic:

> “Candidate is on topic as she states that she has been to a tall building. However limitations in vocabulary mean that she fails to provide sufficient examples or explanations” (R09)

> “She tells us that she visited a very tall building but gives very little information about the experience. I am left wondering where, when, why etc. I still decided it was on topic as she did mentions (sic) the words 'very tall building' at the start.” (R35)

These perceptions are examples of raters who marked the response borderline:

> “The candidate talks about an un-named tall building.” (R07)

> “She talked about a tall buidling (sic) but didn't explain where it was” (R24)

> “Says she visited a tall building but gives few details” (R31)

Rater 36 marked the response off-topic and appears to perceive that the response is fully missed:

> “Does not deal with the prompt.” (R36)

However, not all raters decided the response lacked detail:

> “According to me this is on topic as the speaker describes the building - old, beautiful, crowded.” (R13)

> “Speaker directly addresses the prompt, describing what happened in some detail.” (R16)

> “refers to it by age and beauty” (R22)

For prompt 2, there were 27 on-topic decisions and two borderline and one off-topic. The dominant vocabulary in the perceptions was ‘exciting’, ‘liked’ and ‘afraid’:

> “Exciting /liked it very much/Afraid” (R08)

> “Exciting, liked it a lot - building was old and beautiful. Was a little afraid of tall buildings” (R27)

One rater mentioned adjectives as a reason for an on-topic decision:

> “Speaker used a few adjectives to describe how she felt.”

The two raters who marked borderline mentioned the response being vague or too short:

> “Seems a bit vague on this.” (R07)

> “While there is some description of feelings, this is very short. (R13)

The rater who marked this response off-topic simply commented:

> “fails to comment on this prompt”. (R09)

Prompt 3 was mostly rated on-topic (19 raters) but there were 10 raters who marked it borderline and one who rated it off-topic. For on-topic perceptions, the predominant evidence was that it is ‘important’ and ‘people enjoy’.

> “She says it's important, and that people enjoy them.” (R35)

The raters who did not mark on-topic felt the language level was insufficient:

> “some comment is given by the candidate in an attempt to respond to this prompt, but limitations in vocabulary result in an insuffcient (sic) response to call it 'on topic' (R09)

> “says it’s important to have tall buildings and many people enjoy them; rather vague, but more a problem with language than relevance, I think.” (R06)

> “Low language level, not really sufficient to reach the target level but gives a few phrases such as 'people enjoy it'” (R26)

One rater who marked on-topic noted that the response was not responded to in the correct order:

> “Although the response was not given in the right order and was somehow included in her response to question 1.” (R30)

The three raters (R27, R21 and R34) who scored 4 (B2.2) note the various issues (e.g., no mention of building name, brief, language) but believe the response is at B2 level and score all three responses on-topic.

V1-155: Highest Aptis score and Aptis misfit

This response had the highest fair average Aptis score and was also misfitting in Aptis decisions. Rater 08 had an unexpected score of 3 for this test-taker response.

For prompt 1, all raters except for one rated the response on-topic. This perception explains the reasoning for on-topic decisions:

> “The speaker speaks about practising Tai Chi as a child. He talks about the connection with the sport throug (sic) his mother who was Chinese but living in an Italian town. He lists the benefits - health, made new friends.” (R13)

Other raters had shorter perceptions, for example:

> “practising Tai Chi” (R08)

The one rater who marked this response as borderline wrote:

> “He referred to Tai chi as an activity he used to do as a kid. He mentioned this was usually done outdoors, but it doesn't seem Tai chi is necessarily an outdoor activity.” (R30)

Rater 30 was not able to give an Aptis rating and instead wrote: “I'd escalate this.” This issue of activities not considered outdoor activities was raised in the qualitative analysis (section 4.2.1.2.1). Activities including calligraphy, bike stunts and table tennis were activities some raters did not consider to be outdoor activities.

Except for rater 30, raters perceived that the response was on-topic and had answered the prompt 1: *Tell me about an outdoor activity you did as a child*.

Prompt 2 was also mostly rated on-topic. There was a range of vocabulary used to decide that the response was on-topic such as strange, enjoyed, learn (another

culture), discipline, (physical and social) benefits, made friends, ethnical/ethnicity and socialising. This response was typical:

> “He specifically says he 'really enjoyed it' also talks about some of the benefits he got from it.” (R35)

Rater 10 gave a fuller perception with quotes from the response:

> “I would go there with my family, my brother and two of my cousins. It was outdoor.. the best part of it when the weather was good, I really enjoyed it. I was learning discipline, having a bit of fun with my cousins, I think I really had a good time....that really helped build immune systems, make new friends, in a way know another culture as well.” (R10)

Two raters marked this response borderline with perceptions:

> “Enjoyed it, learning discipline” (R27)

> “strange/ethnical” (R15)

Rater 33 marked it off-topic with the comment:

> “TT did not answer this prompt.” (R33)

Prompt 3 had a range of perceptions. Eleven raters decided it was on-topic, fourteen raters chose borderline and five selected off-topic. Rater 22 highlights the dilemma:

> “implied because of all that he learned but not specifically said”. (R22)

This response was the highest rated, but with a misfit. There were two scores of 6 (C2), 14 scores of 5 (C1), seven scores of 4 (B2.2) and six scores of 3 (B2.1). Raters clearly rated this response at a high level, but the six raters (R08, R17, R20, R21,

R25 and R33) who scored a 3 (B2.1) appear to have penalised the test-taker for topic relevance. Raters 8, 17 and 21 marked prompt 3 off-topic and wrote:

> “not addressed -just talked about himself” (R08)

> “TT describes their own experience and personal benefits from practising tai-chi which does not answer Q3.” (R17)

> “Too much time spent o [sic] his own experiences, not enough time to get fully on-topic with the final answer.” (R21)

Raters 20, 25 and 33 scored prompt 3 borderline and noted that the test-taker:

> “does not complete his reponse (sic).” (R20)

> “says why important for him but not others“ (R25)

> “did not extrapolate to the broader population.” (R33)

The other raters who scored prompt 3 borderline or off-topic, ignored their perception and scored 4 (B2.2) or higher, for example:

> “Personalised this answer in refernece (sic) to his own experience and did not have enough time to answer the question.” (R34)

> “Mainly speaks about personal experience only; at very end starts to speak more generally but doesn't answer prompt” (R31)

Or inferred their decision:

> “Not directly addressed but tangentially addressed in other responses” (R29)

An example of an on-topic decision for prompt 3 is exemplified by rater 28:

> “Some reading between the lines is needed, nevertheless, this is on topic, because the TT talks about the need for learning discipline & learning various things around that age, implying that these type of activities are helpful for that” (R28)

Raters appear to be conflicted on if referring to oneself in prompt 3 should be rated on-topic, borderline or off-topic. The prompt is meant to elicit a response about wider society (Is it important for children to do outdoor activities?). In this response, the response is at a high CEFR level for language but one of the prompts is not correctly answered with respect to the personal. As mentioned earlier, one rater (R30) did not rate the response and wrote that they would escalate to a senior rater, providing further evidence of a difficult response to rate.

V1-119: Aptis and on-topic misfits

The Aptis scoring ranges from B1.1 (eight raters), B1.2 (3 raters), B2.1 (10 raters), B2.2 (3 raters), C1 (6 raters) and significantly misfits. Unexpected score for rater 19. Prompt 1 is misfitting (outfits).

For prompt 1, rater perceptions identify three areas where raters diverge. The test-taker does not speak about themselves, they do not talk about childhood, and they focus on describing the photo. Most raters (23 raters) marked this response off-topic, while four raters marked it borderline on-topic. Three raters marked it on-topic.

Raters who marked it off-topic either commented: “The speaker does not address the question” (R14) or “Describes the activity in the picture” (R15).

Raters who decided the response was borderline note that the test-taker does not reference their personal experience as a child, for example: “makes reference to an

activity, but not directly applied to himself." (R09) and "Doesn't mention anything about TT's personal experience during his childhood" (R12).

Comments from the raters who marked it on-topic include: "cycling was the best activity I guess" (R08) and "The candidate used the photograph as a prompt to talk about cycling." (R07). Rater 20 writes: "talks about cycling in general not as an activity he did" but decides the response is on-topic.

In prompt 2, raters noted that because the first response was off topic, it was impossible to answer prompt 2. No rater marked this response on-topic. Most raters marked this response borderline (24 raters), for example: "Same as the first prompt TT fails to anwer (sic) the 2nd prompt, speaks about benefits of outdoor activity" (R12). Rater 29, who also scored borderline wrote: "Inference from response to third prompt – tangential". Rater 35 noted the lack of speaking about themselves: "He talks a lot about how great cycling is for people but doesn't personalise at all. Its (sic) hard to be clear if he's saying HE enjoyed it."

For prompt 3, all raters marked it on-topic. Raters agreed that the test-taker really only answered this prompt. They noted that the response was detailed and clear and included references to grammar, vocabulary and fluency for example rater 18: "He only answers this question. Excellent range of complex grammatical structures, cohesion and lexis." Rater 14 sums up many other rater perceptions: "I find these high-level, partially relevant, mostly incomplete responses really difficult to mark! I'm never sure what is expected of me."

The high level of language but with topic relevance concerns for two of the prompts is the reason for the misfit for the Aptis scores. The scoring ranges from 1 (eight

raters), 2 (3 raters), 3 (10 raters), 4 (3 raters), 5 (6 raters). It is probably that most raters would escalate this type of response for senior marking.

V2-145: Aptis and on-topic misfits and severe on-topic decisions

Lowest on-topic scores (combined). This response was significantly misfitting for Aptis scoring. Unexpected score for raters 08 and 21.

For prompt 1, this response was mostly rated off-topic (23 raters) or borderline on-topic (7 raters). No rater marked this on-topic. The issues included not addressing the prompt and not speaking about themselves.

Prompt 2 is also rated mostly off-topic (11 raters) or ambiguous/borderline on-topic (18 raters). The main issue is that the test-taker did not speak about themselves. One rater (R06) rated this response on-topic with a comment: “tell [sic] all the advantages and benefits of travel rather than specifically saying how she feels about it”. This responses demonstrates the difference between raters who think it is important to speak about the personal versus other raters who do not.

Prompt 3 is mostly rated on-topic (19 raters) but some raters marked it off-topic (4 raters) and some rated it ambiguously/borderline on-topic (7 raters). An example of an ambiguous/borderline rater perception includes: “The speaker talks in general terms rather than in connection with her own experience or in connection with the experience of people in her country. However, she is on topic in that she is talking about long-distance travel. I would have to mark her down though for not answering the question exactly.” The test-taker speaks generally about travel and is not specific about modes of transport in their country. Raters also noted that the response is not linear, in that the test-taker does not answer each question in order (this was a key

finding in the pilot studies). It could be that because the response is not linear, the response is more difficult to mark for topic relevance.

The Aptis scores ranged from zero (below B1) (four raters), 1 (B1.1) (6 raters), 2 (B1.2) (3 raters), 3 (B2.1) (10 raters), 4 (B2.2 (3 raters), 5 (C1) (3 raters) and one 'I'm not sure'. One rater noted that the language is at C1 level (and therefore gave it a 5) but that the responses are off-topic (prompt 1) and ambiguously/borderline on-topic (prompts 2 and 3). This test-taker response was difficult to mark and would likely be escalated in live marking. However, the range of Aptis scores that raters chose to give demonstrates how they perceive the importance of the on-topic rating requirements. In some cases, with scores of 0, 1 and 2, topic relevance is more important to the rater than the language. Tahe range of views also explains the misfitting Aptis scoring.

V2-120: Highest on-topic score

This response had the highest number of on-topic decisions and also had the second highest Aptis score.

For prompt 1, key vocabulary for a journey (e.g., car, driving, visiting places) is mentioned as evidence. Raters also noted that the test-taker goes into detail and speaks at length. One rater noted that the test-taker used the words from the prompt (R09). All raters marked this response on-topic.

For prompt 2, raters focused on adjectives such as 'thrilling' and 'awesome'. They mentioned 'good memories' and a 'nice experience'. Eight examiners were not specific but marked on-topic. Rater 15 marked the response on-topic with the comment "thrilling experience (indirectly)", which is implicit evidence.

Two raters marked this response as borderline. Rater 26 had the comment: “Describes briefly as a thrilling experience, awesome and nice, but not directly relating it to the question”. Rater 13 had the comment: “The speaker called the experience "nice" and "aweome" [sic] on two different ocassions [sic]. He also said he felt it was thrilling - these may not explicitly answer the feelings question, I would still consider them valid. Was not sure so marked ambiguous instead of on-topic”. However, these raters did not penalise the test-taker in the Aptis score. Both gave the test-taker a score of 4, which means that all three responses are on-topic.

For prompt 3, all raters marked the response on-topic and mentioned train travel and rail network. Raters mentioned that it was a detailed argument. One rater queried: “Did he mention method of transport AND state it was for long distances?” (R19) but marked it on-topic and gave a 5 for the Aptis score (C1 level).

V2-138: Highest on-topic score

Raters gave Aptis scores with an observed and fair average of 4.17. Raters 20, 35 and 36 rated it a 3 and rater 09 rated it a 2. As all raters, except rater 34 rated this response as fully on-topic for all three responses, these four raters are scoring at B1 level for some of the responses. The rest of the raters were split between a score of 4 (B2.2) (16 raters) and 5 (C1) (9 raters).

For test-taker response V2-138, all raters marked prompt 1 on-topic. Raters mentioned “bus”, “8 hours” for length of journey and “school trip” as their evidence, and also noted that it was a detailed response.

For prompt 2, one rater marked it as borderline (R34) with the comment: “Mentioned it was tiring but enjoyed the trip. Perhaps quite a short answer.” However, the rater gave this test-taker an Aptis score of 5 (C1).

For prompt 3, raters had a more varied approach. Some raters mentioned only “cars”, others mentioned “cars”, “buses” and “planes”. Raters mentioned that the test-taker gave reasons, expanded, elaborated, mentioned the prompt, talked about the pros and cons and gave a detailed response. All raters chose on-topic for this test-taker response.

5 Discussion

This study focused on the understudied construct of topic relevance and the internal criteria raters report using to decide if a speaking response is on-topic. We do not know enough about how and why raters make topic relevant decisions and variance in this decision could lead to inconsistent scoring. The Aptis rating scale has a specific yes/no binary topic relevance decision and provided an opportunity to study this subject.

The goals of this thesis outlined in the introduction chapter (section 1.1) were to identify features in the speaking tasks used in the current study which may affect topic relevance decisions; start to develop a construct definition of topic relevance for the Aptis task used in this thesis; contribute to the theoretical understanding of the topic relevance construct in speaking tests more generally; and develop better rater guidance and support materials that explain the speaking topic relevance construct to raters. To achieve these goals, this thesis addressed the following research questions:

1. What internal criteria do raters report using to decide if a speaking response is on-topic?

    1.1. What generic criteria do raters report using that are common across all speaking test prompts and versions used in this research?

    1.2. How do B2-level Aptis speaking task prompt types affect raters' decision-making?

    1.3. How do the task versions from different forms of the test affect raters' decision-making?

1.4. How do the internal criteria that raters reportedly use reflect the criteria in the rating scale descriptors?

2. What is the relationship between the internal criteria raters reportedly use to make their on-topic decisions and raters' Aptis scoring and on-topic decisions with respect to rating severity, reliability, consistency and construct relevance?

The main findings in this investigation discovered that there are internal criteria that some raters use across all tasks and versions to make topic relevance decisions (RQ 1.1), for example the expectation of topic development (quantity and quality of a test-taker's response). In addition, the words and photos in the prompts (RQ 1.2) and the themes in the versions (RQ 1.3) affected raters' topic relevance decisions. There were 10 different elements which raters used to make their topic relevance decisions in this study. Raters were on a spectrum of only using one element (e.g., the theme of the question prompt) or all 10 elements when making their decisions. This thesis also found five approaches rater's use to document evidence in their perceptions (e.g., a test-taker quote); some raters who had consistent approaches to rating used the same approach to collect evidence for each decision and other raters used different approaches inconsistently or used different approaches for different prompts or versions.

Raters gave evidence that they try to follow the Aptis two-step rating process and rating scale criteria (RQ 1.4). However, some raters noted that it can be difficult to rate all aspects of the performance with only one listen and that they consider their final score based on a 'best fit' decision, rather than a strict two-step decision.

This task does not have a clear construct definition of what it means to speak on-topic. Bringing together the full data set begins to identify elements of a construct

definition (RQ 2). There is evidence of raters who follow the two-step rating construct and use their lenient or severe on-topic decisions to make their on-topic decisions and then use their on-topic decisions to decide their Aptis score for this task (see sections 4.3.1.1 and 4.3.1.2 for examples). There are also raters who do not use their topic relevant decisions to make their Aptis score decision (section 4.3.1.3).

Using a mixed methods research approach, I looked at the research problem through multiple rater lenses: on-topic decisions (QUAN), Aptis scoring (QUAN), on-topic perceptions (QUAL) and self-reported rating procedures (qual). The on-topic decisions and Aptis scoring quantitative data identified raters who were severe or lenient, misfitting or who had bias towards test-taker responses, question prompt types or different versions (forms) of the task. The on-topic perceptions identified raters' self-reported levels of severity, reliability and construct relevance in their decision-making. The mixed methods triangulation integrated the quantitative and qualitative data into a final interpretation. The email exchanges, which included the self-reported rating procedures, clarified the findings in the mixed methods interpretation.

The study involved 30 raters from the pool of Aptis examiners. The raters marked 45 carefully chosen Aptis test-taker B2-level spoken responses to a B2-level task. They provided on-topic decisions to the three prompts in each task as well as the reasons for their on-topic decisions. They also provided Aptis scores for each response. I followed up with semi-structured email exchanges to clarify the data, understand outliers and investigate rating procedures. I used Many-Facet Rasch Measurement to analyse the quantitative data in a fully crossed design. I used comparative thematic analysis to analyse the on-topic perceptions.

I have contributed to knowledge in wider terms by advancing an understanding of topic relevance as a construct and specifically if a yes/no binary topic relevance decision for measuring spoken performance is a valid measure of spoken performance. I have also included areas to expand the socio-cognitive framework.

I approached this research in a similar manner to other researchers investigating understudied speaking constructs and rater perceptions. For example, Burton (2020) studied how raters decide whether test-takers are authentically engaged with a test task, an area tangentially related to speaking on-topic. The two studies are similar in their attempt to isolate the non-linguistic aspect of spoken production (authenticity and topic relevance) and to understand if raters can measure or even detect authenticity and topic relevance. However, the data collection methods are different. Burton (2020) used a performance-evaluative approach and semantic differentials across different elements of authenticity (scriptedness, relevance, naturalness, spontaneity, genuineness) on 7-point Likert scales. In contrast, I used a three-point scale (on-topic, borderline, off-topic) to target the rater topic relevance decisions to each of the question prompts. Both approaches were able to isolate the non-linguistic decisions in a meaningful way. Burton (2020) recommends an extended focus on topic relevance and a useful follow-up study would be to identify topic relevance elements and repeat Burton's study. My thesis has identified which topic relevance elements could be included in this type of study, which is discussed in section 5.5 (Future research).

This whole area of topic relevance rating has been underexplored and this thesis addresses this gap in knowledge with a novel contribution. The next sections discuss these contributions in more detail.

5.1 Rater internal criteria to decide if a speaking response is on-topic

This section consolidates findings on the internal criteria that raters report using to decide if speaking responses are on-topic (RQ 1). Included are the sub-research questions regarding decisions that are common across all prompts and versions (RQ1.1), that are specific to task prompt types and task versions (RQ 1.2 and RQ 1.3) and how topic relevance decisions reflect the criteria in the rating scale descriptors (RQ 1.4). This section also discusses the relationship between rater on-topic perceptions (QUAL) and Aptis scoring and topic relevance decisions (QUAN) (RQ 2).

The key internal criteria findings raters use to decide if a speaking response is on-topic are: 1) amount and clarity of topic development; (2) adherence to content in the question prompts; 3) the use of vocabulary; and 4) and linguistic characteristics of the spoken response. There are also five approaches raters use to collect evidence.

5.1.1 Topic development

The first finding was that raters reported using topic development, or the amount and clarity of spoken performance in the test-taker responses, to make topic relevance decisions. This finding was across all prompts and versions (RQ 1.1). Raters had a wide range of expectations on how much detail or elaboration test-takers needed to produce for an on-topic response. Lenient raters wrote short one-word perceptions for deciding if a response was on-topic, for example “cricket” (R08) or “New York” (R15), or used inference to make decisions (“Inference more than actually directly answering the question” R29). More severe raters used question words (who, what, where, when and how) to identify what was missing (“Describes a journey – where, who with, what happened” R31) and checked that all aspects of the question

prompts were answered. As noted in the literature review, topic development is often not well-defined in rating scales and is a difficult criterion to standardise across raters. Therefore, this finding is not surprising and supports and extends the findings of other researchers who also found that raters have different expectations of topic development (e.g., Brown, 2005; Khabbazbashi, 2021; Lee, 2005). See section 4.2.3.1 for frequency counts and a discussion on the amount and clarity of response raters expect.

I did not anticipate such a wide variance on topic development by raters because topic development is not a criterion in the Aptis rating scale. I therefore did not interrogate the raters specifically about their reasons for topic development expectations in the email exchanges. However, there are at least three possible reasons for this topic development finding, which I discuss below. In the future research (section 5.5), I discuss a possible research project to address these three points.

One explanation for different topic development decisions came out of my previous research, which found evidence that IELTS raters struggle with using Aptis rating criteria during training (Fairbairn, 2015a). I concluded that IELTS raters found it difficult to internalise the Aptis rating scale and brought another rating construct to Aptis rater training. At the time, I thought the issue was the analytic nature of IELTS marking. IELTS raters give scores for four different linguistic criteria. Changing to holistic rating and arbitrating between the different criteria to produce one score for all linguistic criteria was the reason for the challenge. There could also be an additional issue that did not occur to me at the time. Marking topic development uses a different approach in the Aptis and IELTS tests. In Aptis, there is no expectation to

develop the topic, while in the IELTS Speaking test, topic development is included in the ‘Fluency and coherence’ scale at the highest levels (levels 8 and 9). Moreover, in TOEFL iBT, not part of my previous research, topic development is an explicit criterion with a rating scale for all levels of performance. Moving between very different approaches to marking topic relevance could be affecting rater decision-making.

Another possible reason why raters may have different expectations of topic development in the context of the current study could be the teaching background variable in the raters. All raters in the study had at least four years of English language teaching experience. The impact of a teaching background on rating is under-researched (Knoch et al., 2021) but it would not be uncommon for teachers to use question prompts that are similar to the ones used in this research to elicit topic development. Teachers would encourage students to tell a story, describe an event or activity or talk about a societal phenomenon. It could be that a teaching background influenced some raters to expect elaboration and expansion, even though topic development is not mentioned in the Aptis rating scale.

A third possible justification for different perceptions could be that some raters expected more content from the easier prompts. The three prompts in this research are meant to be increasingly complex to stretch the test-taker to a B2.2 level. As noted in the literature review, tasks with high levels of complexity and abstractness target higher level performance. However, there could be a type of affective schemata impact, in that raters could have a positive or negative emotional feeling or reaction about a topic, similar to affective schemata experienced by test-takers as noted by Bachman and Palmer (1996). Affective schemata could have influenced

some raters to expect more or less topic development depending on the prompt. Hamp-Lyons & Mathias (1994) and O'Sullivan (2012) both noted that raters can have a reaction to a topic. Some raters may have found the abstract prompts difficult for test-takers and may have been more lenient with their topic relevance perceptions and scoring. It was not possible to tell in the qualitative data if raters considered the prompts as three equal questions, or if they expected more content for easier or more difficult prompts; they did not report on this issue.

Topic development is currently not explicitly part of the construct for the task used in this thesis. However, raters clearly had different topic development expectations and the wide variance of perceptions affected on-topic decisions, which in turn affected scoring (see sections 4.3.1.1 and 4.3.1.2: lenient and severe raters). This finding stresses the importance for test developers to define topic development in the rating scales and train raters on this criterion. It would be unfair to expect topic development by test-takers unless the test developers make this criterion explicit in the question prompts, rating scale, preparation materials for test-takers and training materials for raters.

5.1.2 Question prompt content

The second finding was that some raters focused on the content of the task question prompts to make their topic relevance decisions. These raters focused on task fulfilment and used the words and photos in the tasks to explain why they decided the response was on-topic, borderline or off-topic. There were three sub-findings in this area: grammar expectations; the impact of the picture; and variance in understanding what a feeling means.

5.1.2.1 Grammar expectations

The first sub-finding is that raters had different perceptions on using the grammar requirements in the question prompts to make their on-topic decisions. Some raters considered the response off-topic if there were grammar errors with respect to the question prompt, while others did not. See section 4.2.3.2 for more information. The variance of perceptions affected on-topic decisions and Aptis scoring in this research.

There were three types of grammar decisions that affected raters' topic relevance decisions. The first was the verb tense. There are two different elicited verb tenses in the question prompts (past and present) and some raters reviewed the verb tense accuracy with respect to the question prompts when making their on-topic decisions. The variance of perceptions was particularly noticeable for prompt 2, (How did you feel about the event or activity in the past?). Raters had different perceptions if the test-taker did not speak about their feelings at the time of the event or activity in the past tense and instead spoke about the outcome of the event or activity or about their feelings in the present test. Some raters penalised test-takers for speaking in the incorrect verb tense.

The second grammar decision some raters used to make their topic relevance decision was regarding the task requirement in question prompt 1 to speak about only one singular activity or event. Raters had different on-topic perceptions if the test-taker spoke about more than one activity or event. This finding is linked to the topic development finding. Raters who expected topic development did not accept lists of activities or events; raters who did not expect topic development were not so concerned.

The third grammar decision was about the personal and general requirements of the question prompts. The tasks elicit spoken production about oneself and about wider society. The literature review discussed how speaking about oneself is a lower level of spoken production than speaking about wider society (Council of Europe, 2001; Taylor & Galaczi, 2011). The purpose of eliciting spoken production about wider society in the third prompt is to stretch the test taker to a B2.2 level of spoken performance. Raters had different perceptions if test-takers incorrectly spoke about themselves when they should have been speaking about wider society and vice versa.

Grammar requirements in question prompts are a type of task manipulation to help ensure the question prompts elicit the required level of spoken production (Robinson, 2001, 2015) and are a key part of the task construct. There is a wealth of research on how raters mark test-taker grammar with a focus on accuracy and complexity of grammar production in spoken performance. Bachman and Palmer (1996, pp. 53-54) touch on the grammar expectations of the response by including grammatical characteristics as part of the expected output but do not link grammar explicitly to topic relevance. Brown (2000) found that all raters made comments about incorrect verb tenses, although it is not clear if any errors occurred due to the question prompts. Sasayama and Norris (2022) showed how manipulating speaking task complexity changes test-taker performance. There is no research that I could find which specifically investigates how grammar errors with respect to the question prompts affect raters' topic relevance decisions. This research shows that task complexity manipulations not only affect test-takers' spoken performance but also some raters use the expected grammar complexity requirements in the question prompts to make topic relevance decisions.

What became apparent in this research is that there are two grammar decisions that may be interacting in unexpected ways. The first decision is regarding the overall accuracy and complexity of grammatical structures as part of their CEFR decision. The second grammar decision is the accuracy of the grammar with respect to the question prompt. The evidence shows that some raters did not take the grammar requirements of the question prompts into consideration when making topic relevance decisions because they either did not identify the errors, considered the errors part of the CEFR decision instead of the on-topic decision, and/or did not consider incorrect grammar with respect to the question prompts as an on-topic error. Conversely, it is possible that some raters may have penalised the test-taker twice, once in the overall language decision and then again in the on-topic decision.

This study was able to identify that raters have different on-topic perceptions and decisions with respect to the grammar requirements in the question prompts, which does affect scoring. However, it was not able to identify if raters are making the same or different decisions in the CEFR and on-topic decisions.

5.1.2.2 Pictures

The second sub-finding identified that the pictures in the question prompts can affect raters' topic relevance perceptions. Some raters noted that test-takers talked about the picture rather than answering the question and marked them off-topic as a result (e.g., "TT does not answer the question. They only refer to the picture saying that cycling is the best outdoor activity." R17). Other raters considered a description of the photo as speaking on-topic. Pictures in two of the three task versions used in this research affected topic relevance decisions. The task with a more unusual photo

(crossing a desert by camel) did not affect raters perceptions. See section 4.2.3.1 for the reporting on how the picture performed in each task version.

There is research on picture prompt tasks (e.g., Sasayama et al., 2022) where test-takers are expected to describe the picture but I found no research on tasks that have a picture but test-takers are not expected to describe the photo. My research provides evidence that if a photo is not carefully chosen in this type of task, it can confuse the test-taker and the rater.

As this finding became apparent early in the project, I made a policy change at the British Council on the specifications for choosing photos. There is now a requirement for photos in this task to be abstract. For example, in a recent new task, there is a cartoon picture of a person rolling a huge boulder up a mountain to demonstrate the theme of achieving dreams. Raters have provided feedback that they prefer these images as there are no longer test-taker responses that only talk about the photo. I reflected on removing the photo completely to avoid this confusion. I decided that the abstract photos may provide ideas for abstract concepts and imagination that can help test-takers so I decided to keep the photo.

5.1.2.3 Feelings

The third sub-finding was that some raters struggled with topic relevance decisions for the second prompt about feelings. This prompt is in all three versions in question prompt 2: ‘How did you feel about (the activity or event in the past)?’. The test-taker is meant to reflect on their response to question prompt 1. Some raters did not seem to know how to identify if a test-taker had expressed a feeling, there were some incorrect perceptions of feelings and some correct feelings were discounted, for example: “Doesn’t talk about feelings but describes how it was hard and overcame

the obstacle and learnt from it." (R35). There was some lack of understanding that there are two definitions of feelings: (1) to experience something physical or emotional; (2) and to have a particular opinion about or attitude towards something. See section 4.2.3.3 for a discussion on 'feelings'.

There were comments for prompt 2 (How did you feel about it) that raters felt the response was borderline but they did not penalise the test-taker, providing evidence that test-takers struggle with this prompt and raters may not prioritise this prompt being on-topic when allocating their Aptis score.

There is no research that I could find on rating this type of question prompt with respect to speaking on-topic. However, in the CEFR Companion Manual (Council of Europe, 2018), there is a new *Sustained Monologue: Describing Experience* scale with a high B1 level descriptor "Can clearly express feelings about something experienced and give reasons to explain those feelings" (p. 70), so perhaps research will emerge. Of note is that for the Aptis test, describing feelings is considered a B2-level spoken production, whereas it is a B1 level of performance in the new CEFR descriptor.

5.1.3 Vocabulary

The third finding was that raters identified vocabulary in the test-takers' responses which addresses the question prompts to make their topic relevance decisions. Most examiners used this approach. As my second coder mentioned: "Regarding vocabulary – for me, vocabulary is at the heart of whether a response is on or off topic." (email exchange).

This finding supports the literature that vocabulary is closely linked to topic relevance decisions (Lumley, 2005; Milanovic et al., 1996; North, 2003; Sasayama & Norris, 2022). North (2003), for example, encouraged vocabulary precision in rating scales with respect to topic relevance criteria but warned that without including appropriateness or coherence of the response, using only vocabulary could focus too heavily on memory. My thesis highlights this dilemma. Test-takers could mention vocabulary that is on-topic but which lacked appropriateness and coherence in the response. Some raters scored these test-taker responses on-topic, while other raters did not.

Cai (2015) is one researcher who separated vocabulary and topic relevance. This research had two rating scales, form-related and content-related. The form-related rating scale included vocabulary, while the content-related rating scale included topic relevance, richness in content and organisation. However, Cai (2015) found that some raters tended to confuse content-related criteria with vocabulary criteria. For example, one rater in this research perceived using synonyms as a content decision rather than a vocabulary (form-related) decision. The findings from my research show that separating vocabulary from content is not consistently achieved (see the discussion on ‘test-taker words (examples, quotes and vocabulary) in section 4.2.3.1 for a fuller discussion on how my second coder and I differed in our approaches.

Raters in this research with high construct relevant decisions for topic relevance focused on examples and quotes using vocabulary from the test-taker’s response that responded explicitly to the task requirements. For example, see rater 08 in section 4.3.1.1 (Lenient raters). This rater makes the same type of decision every time, mentioning the key information using quotes in short perceptions.

5.1.4 Linguistic characteristics

The fourth finding was that some raters used other parts of language to make their on-topic decisions such as fluency and coherence. The perceptions also included clear and clarity (plus antonyms and derivatives), usually linked to fluency and pronunciation. See section 4.2.3.1 for the results.

Raters used the words ‘clear’, ‘unclear’ and ‘clearly’ to refer to both speech clarity and clarity of content. The speech clarity type of perceptions suggest the rater is referring to the comprehensibility, intelligibility or understandability of the spoken response, for example “Speech is clear and easy to understand” (R09). This first type of perception was often used in perceptions to discuss fluency, including the speed of spoken production and pausing. Speech clarity is assessed in the CEFR decision so this finding that some raters report using speech clarity to decide if a response is on-topic is an important finding, and possibly construct irrelevant.

Topic relevance is meant to be the non-linguistic part of scoring spoken performance. This split between linguistic performance (CEFR) and topic relevance decisions is explicit in the Aptis test. However, this thesis shows that some raters report using linguistic performance to make their on-topic decisions. As mentioned earlier, it could be that some raters are using the construct in other rating scales that merge topic relevance and other parts of language (e.g., IELTS and coherence). Raters who expect topic development, for example, might expect coherence. Without a clear construct definition, the B2 task in Aptis does not consistently test coherence and higher levels of spoken production.

5.1.5 Approaches to collecting evidence

The fifth internal criteria are the approaches to collecting evidence. I identified five approaches that raters used to collect on-topic evidence. Some raters always used the same approach, while other raters used a few different approaches.

No specific evidence

There were perceptions that did not provide very much information on how and why raters made decisions (e.g., “The speaker gives a relevant response” R14). It is unclear if these raters had internal criteria in mind but did not, or could not, explain their reasons, as cautioned by Dörnyei, 2007. Or that their internal criteria is very basic and they are only looking for a response that is ‘relevant’, ‘appropriate’ or ‘acceptable’. As this is a yes/no decision, they perhaps do not have very much internal criteria to make this decision.

Short quotes

Most raters used one word or short phrases to capture words that directly answer the main parts of the prompt, for example “Volleyball” (R26). This approach to collecting evidence was realistic and practical. It was used primarily by raters who were lenient with their on-topic decisions, on-topic perceptions and Aptis scores. (see section 4.3.1.1 Lenient raters for evidence). These types of perceptions focused on the main theme of the prompt and found vocabulary to address this theme. These perceptions did not consider topic development, grammar expectations or picture interference.

Extensive quotes

Some perceptions used extensive quotes to collect evidence. Rater 10 used this approach for all perceptions. This approach puts the evidence in context and makes it clear that the response is on-topic or off-topic; it clearly provides the evidence. However, using extensive quotes seems long quotes are impractical to capture in large scale rating situations such as Aptis unless raters are paid to listen to the response a few times to record a quote.

Words in prompt

A fourth approach raters used was to refer to the words in the prompt. However, raters differed on which words were important. For example some raters perceived 'long' to be important and other perceived 'journey' to be important in the prompt "*Tell me about a time you went on a long journey*" as is exemplified in pilot study 2 (section 3.4.2.3 Pilot study 2 findings). Moreover, some raters perceived words such as 'you' to imply the test-taker must talk about themselves; other raters did not.

Inference

There were perceptions where the rater had inferred or implied what the test-taker meant to say or interpreted what the test-taker said by reformulating what the test-taker has said, for example "We can infer how the TT feels about it (positively), but they don't directly address this prompt" (R28). Inference was notable for prompt 2 about feelings.

5.2 Theoretical contributions to knowledge

This thesis provides evidence to start a discussion on the topic relevance construct definition for the Aptis B2-level speaking task and the use of on/off-topic binary decisions to mark topic relevance for this task. This thesis also expands our understanding of the topic relevance construct more generally. The next sections discuss these contributions and proposes an research agenda.

5.2.1 Aptis speaking on-topic construct

One of the aims of this thesis was to start to develop a definition of the topic relevance construct for the B2-level task in the Aptis speaking test. Currently, there are three questions in the task and raters decide if the spoken response for each question prompt is on-topic or off-topic. It is a yes or no binary decision without any criteria for making this decision. There is no policy on the information that is given to raters on what it means to speak on-topic.

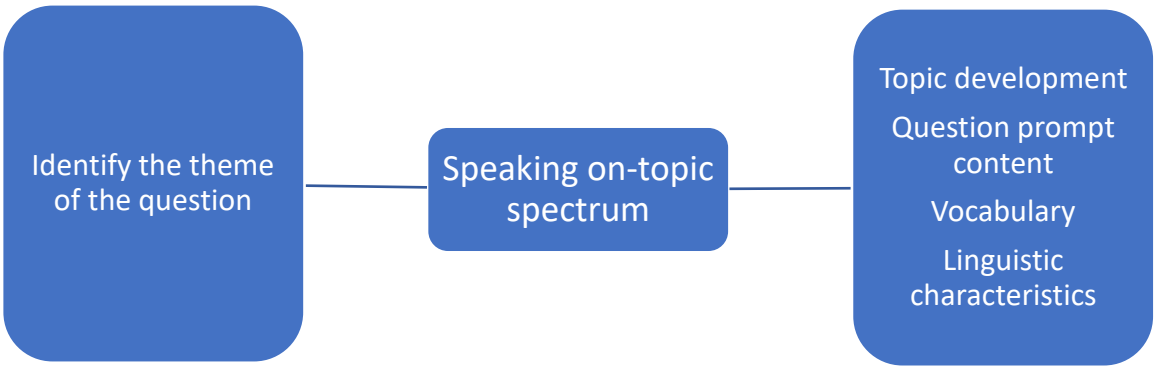
My mixed methods interpretation showed that variance in the on-topic perceptions affects on-topic decisions (see sections 4.3.1.1 and 4.3.1.2 for findings). I have documented examples of raters with lenient on-topic perceptions, on-topic decisions and Aptis scores and vice versa. The B2-level task score is used to determine the score and CEFR level for the Aptis speaking test. The Aptis speaking test is used to calculate the test score and CEFR level for the Aptis 4-skill test (reading, listening, speaking and writing). On-topic decisions, therefore, can affect the scores test-takers receive on the B2 task, speaking test and Aptis test.

With findings in mind from the previous section on the internal criteria raters use to decide if a response is on-topic, the next section discusses possible definitions for an Aptis on-topic speaking construct for the B2-level task used in this thesis.

5.2.1.1 Aptis speaking task construct definition

My research identified that there is a wide spectrum of perceptions by raters of acceptable test-taker topic relevance performance and that these decisions affect scoring. At one extreme there are raters who have short perceptions for topic relevance decisions (e.g., lenient raters in section 4.3.1.1). At the opposite extreme there are raters who require all parts of the question prompts to be addressed in the spoken production for the response to be on-topic (e.g., severe raters in section 4.3.1.2). There is a spectrum of acceptability from one-word perceptions focusing only on the main theme of the prompt to an adherence that all parts of the question prompts are attended to including: grammar expectations; correct use of the picture; topic development using precise and a wide range of vocabulary; and other linguistic demands such as clarity (pronunciation and fluency) and coherence. This spectrum is visualised in Figure 20. There is no consensus by raters on what it means to speak on-topic on this task.

Figure 19: Spectrum of on-topic perceptions



Adherence to all requirements in the speaking prompts can be difficult for test-takers to achieve even if the topic is familiar. As de Jong (2023), Godfroid and Hopp (2023) and Roever and Kasper (2018) explain, the mental processes required to answer

questions during a speaking test is challenging. Perfect and unambiguous spoken communication by test-takers is generally unattainable (Bygate, 1987).

The requirements in an Aptis B2-level task cover at least 10 different on-topic elements across the three question prompts. To specify that all on-topic elements must be adhered to, a rater listening for two minutes to a spoken response, usually only once, would need to correctly identify that a test-taker had satisfied the following criteria: used the correct theme and sub-themes with appropriate vocabulary for each of the three prompts, which includes the ability to identify feelings for prompt 2; used the correct verb tenses (past tense for prompts 1 and 2 and present tense for prompt 3); focused their spoken response only on a singular activity or event for prompt 1 rather than multiple activities or events; discussed a personal experience for prompts 1 and 2 and provided a general comment on society for prompt 3. These decisions would be on top of their CEFR decision, which requires the rater to holistically assess the test-takers' overall grammar, vocabulary, fluency and cohesion.

This research allowed raters to think about each prompt and how and why they would make their on-topic decisions, which gave them an opportunity to tell me what they think is important. In live rating, raters would not have this amount of time to carefully think about their decisions. Requiring strict adherence to all parts of the question prompts, in my opinion, would require raters to listen to responses more than once and for the decisions to be moderated by senior raters. This approach would lead to more severe ratings on the B2 speaking task, the Aptis speaking test and the four-skills Aptis test.

Seedhouse and Harris (2011) discussed this type of broad approach and recommended that raters ask themselves if the questions are answered clearly and coherently with sufficient time spent on topic development. However, they note the difficulties in defining what constitutes a topic and the complexity of systematically investigating adherence to topics. My research has extended these findings by also agreeing with this complexity. Measuring quality is prone to rater bias. Measuring the extent to which the test-takers' response meets the task requirements in the question prompts is difficult to quantify.

On the other extreme, accepting only one-word perceptions as on-topic evidence with a focus on identifying that the test-taker has addressed the main theme of each question prompt is easier for raters to achieve and standardise and therefore would have more reliable scoring. However, accepting one-word perceptions based on the theme of the question prompts could be difficult to explain to raters. There could be a lack of face validity, or credibility, in the rating criteria, in the eyes of the raters.

Hymes (1972) recommends listeners ask themselves if a spoken utterance is possible, feasible, appropriate and performed. This approach focuses less on topic clarity, coherence and development and instead asks the listener to focus on if the response is viable. It could be a suitable approach in a yes/no topic relevance construct definition and centres on the concept of communicative competence, which Hymes (1972) introduced. This approach may be more practical and easier to standardise. However, as Widdowson (1989) noted, there are still degrees of acceptability.

Hymes (1972) was not considering language testing but the four questions struck me as possibly an approach some raters may use to consciously or unconsciously make

topic relevance decisions. I did not specifically ask raters if they used Hymes' questions when making decisions but there could be some implicit use of the questions (e.g., raters describing how a response is feasible). One approach to overcome lack of face validity in a one-word perception approach to making on-topic decisions is to use Hymes' questions within a 'can do' framework. An appropriate response that is communicatively competent should be acceptable if it is possible, feasible, appropriate and performed.

Pallotti (2009) had a similar approach to Hymes (1972) and defined spoken response adequacy as the degree to which a test-taker's performance successfully achieves a task's goals. This approach does not include linguistic characteristics of the spoken response and instead focuses only on task achievement, fulfilment or adequacy, or communicative competence. Khabbazbashi (2021) also showed that a test-taker can speak on-topic without necessarily even knowing anything about a topic. My research also included a test-taker who had not experienced the topic (being in a tall building) and extends this research. Raters need more support in how to decide if these types of test-taker responses are on-topic.

A yes/no decision in a computer-delivered test is possible but only if the criteria for making the decision is limited and focused on communicative competence. If there are degrees of compliance, then a rating scale may be preferable. Khabbazbashi (2021) was able to separate test-takers into 4.7 statistically distinct topic development ability levels so there is evidence that raters can distinguish between levels of topic relevance.

There are difficulties with both extremes in this spectrum. North (2003) warned of the risk in using rating scales with degrees of topic relevance performance but also of

the danger of using binary yes/no decisions. However, defining a middle ground for rating topic relevance is difficult to achieve: what should be included and how can it be measured?

There are nine areas to consider when thinking about this spectrum and developing an on-topic construct for the Aptis B2 task. These include the: purpose of the test; machine interlocutor; length of spoken response; malpractice; grammar for CEFR and on-topic decisions; number of listens; two-step CEFR and on-topic decisions; planning time; and picture.

1. Purpose of the test: The Aptis test is used for a variety of purposes as documented in the introduction chapter of this thesis. The variety of purposes makes it more difficult to define a topic relevance construct. How important is it to test users to measure test-takers' adherence to prompts? Are institutions who use the test for particular purposes such as university admission (high stakes) or placement onto language courses (low stakes) interested in adherence to topic relevance and if so, why? Or are institutions satisfied that speaking test prompts be used primarily to elicit spoken performance?

2. Machine interlocutor: The second area to discuss is the lack of a human interlocutor who can probe the test-taker. Without a live examiner to correct and probe test-takers, is it fair or important to strictly measure topic relevance?

3. Length of spoken response: The two-minute length of the spoken performance gives ample opportunity to develop a topic. However, it is possible for a test-taker to focus minimally on two prompts and then use the rest of the time to speak about one prompt. The prompts target different CEFR levels so the test-taker

could focus primarily on the easier prompt and have one-sentence responses for the other two more difficult prompts.

4. Malpractice: If topics are not expected to be developed, there is a risk that a well-prepared test-taker could minimally answer all three questions and then shoe-horn a prepared tangential response into the spoken performance. If the stakes of the test are high, this is a serious possibility.

5. Grammar for CEFR and on-topic decisions: It was shown in this research that raters may use their grammar decisions as part of their topic relevance decisions, CEFR decisions, both or neither. These included the expected verb tense and speaking of themselves or wider society (personal or general) and speaking about singular activities, journeys or buildings, rather than multiple. This discussion would need to agree where grammar decisions take place.

6. Number of listens: It would be difficult for raters to accurately judge all on-topic decisions when making their decisions if only listening to a response once. There are cost implications to listening to responses multiple times so how many times raters are paid to listen to a response is a factor in the creation of the topic relevance construct.

7. Two-step CEFR and on-topic decisions: Raters are expected to decide on a CEFR level and then fine-tune the decision using the topic relevance decisions. According to the email exchanges with raters on their marking style, this does not happen all the time. In practice raters are arbitrating between the two decisions as they listen to the test-taker's response. A quote from an examiner describes this issue: "The construct requires a 2-step approach CEFR level, then on-topic. In practice, unless listening twice, these are simultaneous decisions" (R11).

There could be a tendency to pragmatically follow a ‘best fit’ approach and mark holistically across all criteria. There are other raters who ignore their on-topic decisions and collaborate research by other researchers such as Brown (2005), Cai (2015) and Khabbazbashi (2021) who showed that there are raters who mark primarily on language (see section 4.3.1.3 for Mixed profile raters for examples).

8. Planning time: Test-takers have one minute to prepare a response. No rater in this research mentioned the planning time as a reason for their on-topic decisions. Inoue et al. (2021) showed that planning time did not affect test-taker performance in an integrated listening-to-speaking test. Nonetheless, it would be prudent to discuss and research whether planning time changes any of the topic relevance construct decisions.

9. Picture: The visual aid must be carefully considered.

As can be seen, there are a number of areas to consider in the development of an on-topic speaking construct for the Aptis test. This work requires a panel of experts from the Aptis rating team to discuss the points above, agree the construct and then develop a communication plan to roll-out changes to raters.

The work involved in developing an Aptis topic relevance construct could be in two parts: 1) improve the current yes/no on-topic decision-making process for lower stakes tests; and 2) develop a topic relevance rating scale for high stakes Aptis tests.

5.2.1.2 Rater guidance and support materials

Another objective of this thesis was to develop better rater guidance and support materials that explain the speaking topic relevance construct to raters. Researchers

agree that to help ensure rating reliability, rater training and standardisation are required (Alderson et al., 1995; Davis, 2016; Elder et al., 2007; Hamp-Lyons, 1995; Knoch et al., 2007; Lim, 2011; Shaw, 2002; Weigle, 1994, 1998, 2002; Wigglesworth, 1993; Xi & Mollaun, 2011).

Before training can be developed, the Aptis marking management team must come to an agreed understanding of what it means to speak on-topic in the Aptis context. There are 15 people on the marking management team and a workshop to review these results and agree on the Aptis on-topic construct needs to first take place as discussed in the previous section. The nine areas for discussion would be the agenda for the workshop.

Standardisation for existing raters must be carefully managed or there could be a risk that senior raters over-emphasise changes with raters. The management team need to carefully develop a communication plan for the raters, explaining the on-topic construct. Standardisation to help raters mark topic relevance more reliably involves raising awareness about language testing and could include four training activities discussed below.

The first training activity is to raise awareness about different topic frameworks as discussed by Brown and Yule (1983). Some raters might need more support to understand the various approaches to rating the on-topic construct, especially raters who regularly mark other tests. Research studies tend to view previous experience of rating other tests as a positive (e.g., Lumley, 2005) but rater background experience using different types of scales can affect decisions (Fairbairn, 2015a). Raters may not have thought about different topic relevance rating approaches and may not realise they are using another rating construct to make topic relevance

decisions in the Aptis test. Orr (2002) noted that raters often lack understanding about the model of language ability on which the rating scale is based. This type of training about test constructs may help to reduce extreme decisions.

The second training activity could focus on the question prompts and vocabulary. Raters could look at question prompts and identify vocabulary that meets the requirements of the prompt. An awareness raising session that discusses unusual vocabulary that still meets the requirements of the prompt, for example unusual outdoor activities or the difference between a holiday and a journey, could be included so that raters think more deeply about their own biases. Alternatively, raters could review test-taker responses without seeing the prompts and identify the key words in the question prompts, as suggested by Brown and Yule (1983).

The third training activity could bring together a group of raters to code a selection of the test-taker responses from this study and compare coding. It was a revelation for me to compare my perceptions with my second coder. My second coder noted: “To be honest, I had always thought deciding if something was on/off topic was an easy task (that’s how I always felt when I mark), now I see that there are so many nuances and different rater perceptions! Goodness – there is so much to discuss!” (email exchange). Rather than looking at example test-taker responses that are on-topic or off-topic, standardisation could involve looking at rater perceptions and agreeing which perceptions are severe or lenient, reliable and construct relevant.

The fourth training activity could replicate the concept mapping study by Papajohn (2002). Raters could be asked to visually represent their step-by-step rating process using flowcharts. Rater trainers could use the information gained from the concept

mapping to emphasize key rating criteria that raters have deemphasized or overlooked.

There are two training activities that I considered but do not think will work. The first is to review speaking tasks with raters before they go live. During my PhD, I commissioned a small-scale pilot to investigate this idea. The researcher pretested speaking and writing tasks using raters as expert judges. Three experienced Aptis raters reviewed five new speaking and writing tasks using a survey designed by a researcher in collaboration with the Assessment Development team at the British Council (Rossi, 2021). The researcher asked the raters: 1) what would constitute an on-topic response to the tasks; 2) whether there was a possibility of legitimate off-topic responses; 3) if the task would be clear to the test-taker population; 4) if the task was testing the intended construct; 5) if there might be any difficulty marking the responses; 6) the suitability of the pictures selected for the task; and 7) recommendations for task improvements. Findings were similar to my research in that raters had different expectations of what test-takers should cover in a spoken response. The only area of convergence amongst raters was the suitability of the photo. There were very few comments on construct relevance.

It was likely that the project was challenging for the raters because there were no test-taker responses. They had to imagine responses and speculate. We also did not train the raters to be item writers. The recommendation coming from this research was that the senior item writers should do this analysis, not the raters. Aptis has a checklist for senior item writers similar to the one developed by Khabbazbashi (2021) and we should continue with this approach.

The second training idea that would not work is to provide guidance to raters on the key elements of new prompts that they should use to assess topic relevance in test-taker responses. Knoch et al. (2021) recommend separate materials for each task version with annotated sample responses. Realistically, with thousands of questions prompts in our item bank, this would be near impossible to manage.

This thesis has identified that the on-topic construct is not clear and work is needed before rater guidance and support materials can be prepared. However, this section has provided some practical enhancements to the Aptis rater training and standardisation.

5.2.2 Topic relevance construct

This section discusses the findings with respect to the topic relevance construct more widely and recommends that topic relevance decisions by raters could be added to the socio-cognitive framework (Weir, 2005). The current framework explains how to establish scoring validity evidence with respect to topic relevance but does not include raters' perceptions of test-takers' spoken responses. The framework discusses rater reliability in general but does not specifically discuss how raters deal with topic relevance decisions made by test-takers. The focus also is mostly on spoken interaction with an interlocutor. There is scope from my research to widen the socio-cognitive framework to include rater perceptions of topic relevant decisions made by test-takers in a semi-direct speaking test.

The scoring validity chapter in Weir (2005) could include a section regarding the complexity of topic relevance scoring and the effects of different types of topic relevance rating approaches. These approaches should be discussed with their benefits and drawbacks and how they link to the purpose of tests. The examples

Weir (2005) uses in chapter 11 to exemplify speaking test development could include a series of questions to help test developers choose a topic relevance scoring approach. Weir (2005) does not tackle topic relevance rating in the writing section but these proposed enhancements would be equally relevant for writing so there should be stand-alone sections in both chapters regarding topic relevance.

The stand-alone sections should include information to consider for the test specifications about the features of the input and expected responses with respect to topic relevance and rating so that test developers can justify their choice of topic relevance rating approach.

Adapting the areas raised in the Aptis speaking construct definition discussion in the previous section is a useful starting point for discussing the topic relevance construct more generally and are displayed in Table 48. These questions could be used to help testing professionals develop a topic relevance construct for a speaking test or task.

Table 48: Topic relevance construct and the socio-cognitive model

Areas to discuss	**Discussion points**
Purpose of test	What is the purpose of the test? Who are the users of the test? What are the most important parts of language you want to elicit in the test? Is topic relevance important to the users of your test?
Interlocutor	Does your test have a human or computer interlocutor? Is there an opportunity to engage with the test-taker to bring them back on-topic? Is there an opportunity to probe the test-taker to ensure they speak on-topic?

Language criteria	What CEFR level are you testing? Would you like topic relevance to be integrated with other language criteria? Which criteria (e.g., coherence, vocabulary, fluency)? Are grammar decisions part of the topic relevance decisions? What is the order of decisions? Are the decisions equally important? Is there any tension between rating the quality of a test-taker's response and task fulfilment (requirement of the question prompts)? Would you like topic development as a stand-alone requirement? What would you do if a C2 level test-taker spoke off-topic?
Topic development	Do you expect topic development? Why? What would you expect a test-taker to accomplish for a high score in topic development? Do you expect raters to consider coherence when making their topic relevance decisions? What is the minimum number of sentences to a question prompt that is acceptable for topic development? What is more important to you: quality of the response, adherence to the question prompts or accuracy and variety of words? How do you expect raters to collect evidence?
Length of response	How long is the test response? Is there ample time for topic development? Is there an expectation that the time will be used in a particular way (e.g., equal time per question or focus more time on one question?).
Number of listens	Can the rater listen to a test-taker response more than once? Are raters paid to listen more than once?
Malpractice	Are you concerned about memorised responses? Do you expect malpractice? Do you have the ability to multiple mark test-taker responses?

Planning time	Is there any planning time? How long? Does planning time affect rater topic relevance decisions?
Visuals	Do you plan to use visuals (e.g., photos)? Why? What do you expect raters to do with the photos? Is it possible for the test-takers to use the visuals inappropriately? What would you expect raters to do if test-takers use the visuals inappropriately? Is it ok for the test-taker to only describe the photo?

There is no correct way to measure topic relevance. Using these questions could help testing professionals decide on the topic relevance construct suitable for their tests.

5.3 Methodological contributions to knowledge

The global nature of the thesis and for reasons of authenticity, this thesis used write-aloud protocols and email exchanges to collect data. This section discusses these methods and the contribution to knowledge.

Raters were asked to write why and how they made topic relevance decisions. There were no instructions on how to write this information. The method was authentic as some raters mentioned in the email exchanges that they take notes while marking speaking tests to remember salient features and to count how many of the questions are on-topic. Therefore, this method had less disruption on decision-making and rating. As Baker (2012) notes in a similar research project that used write-aloud protocols, speak-aloud protocols would have been difficult to conduct while marking and retrospective spoken reports may not have captured the data fully or authentically. Baker (2012) was studying the writing skill and her argument is even

more correct for the speaking skill. Moreover, the data did not need to be transcribed and interpreted. The words of the raters are exactly as the raters have recorded them.

Baker (2012) provided more information to raters on how to take notes including: their feelings as they made decisions; the approximate time during the reading of the written test-taker response when their decisions were made; their comfort level or level of confidence in their decision; the time it took them to make their decision; what they might have wanted/needed as they made their decision; any other information that they viewed as salient at the decision-making moment. In my research, I did not provide this extensive information and asked only ‘how’ and ‘why’. I followed up with a question about their perceived feelings when making decisions, which probed comfort level and confidence. I did not investigate the approximate time when topic relevance decisions were made as this would have been complicated in a speaking test.

This research also used email exchanges as a method to collect data. It worked well for two reasons. The first reason was because the raters were expert users of English. They were able to articulate their views and interact with the email exchanges. There were no feelings that raters were struggling to communicate with me. Secondly, raters are globally dispersed so communicate with British Council managers about marking issues by email. Therefore, email exchanges were a familiar, authentic and a comfortable data collection method.

5.4 Limitations

The first limitation is that in the pilot studies I collected on-topic decisions and perceptions for the full test-taker response, which was different to the main study. In

the main study, I separated the question prompts and asked raters to provide on-topic decisions and separate perceptions for each question. I noticed one key difference in the perceptions. In the pilot studies, there raters reported perceptions that high level test-takers integrate the three prompts into a long turn response and low-level test-takers answer each question in order. Some raters in the pilot studies also reported that it was more difficult to mark the on-topic rating requirements in an integrated response. This finding highlighted that test-takers could either be penalised or given an advantage for producing an integrated response. The sample of raters in the pilot studies was small so it is unclear whether this finding is significant. By separating the questions, the on-topic decision-making data were richer and better addressed my research questions on rater values about the on-topic rating construct more explicitly, but the decision-making process could be different in live marking where there may be more of a focus on coherence.

The second limitation is that raters could listen to the responses as many times as they wanted in this research study. I wanted raters to think carefully about their decisions and perceptions. In an operational setting, raters do not have the luxury of time to think deeply about only this rating criterion. Some raters might be less able to identify their reasons under time pressure, especially raters who gave long quotes, without listening more than once to the test-taker response.

The third limitation was the amount of time spent on the email exchanges. I could not probe each rater on everything that I wanted to discuss and instead focused on the most obvious findings. As the PhD progressed, I at times wanted to go back to the raters to probe more deeply. For example, I did not notice the ‘no evidence’ perceptions early enough (e.g., “The speaker responds relevantly and well”) and

therefore did not probe in the follow-up email exchange stage of the project on what this means. It would have been prudent to probe these raters for more detail of the 'how' and 'why' they make decisions so I could better understand their reasons.

The fourth limitation is that I did not ask raters why or how they decided on their Aptis scores. I made the decision not to ask them about their Aptis scores so that I could better isolate their topic relevance decisions. I can only assume that the topic relevance decisions are the reason for their Aptis scores but this might not necessarily be true, and in fact I found some raters who do not use their topic relevance decisions (see section 4.3.1.3 Mixed profile raters for examples).

The fifth limitation is the coding. There was ambiguity in some perceptions and only 10% of the data was second coded. The ambiguity did not affect findings but more second coding with discussions between coders regarding ambiguous perceptions,

5.5 Future research

This research was exploratory in nature and has provided me with a solid foundation to pursue on-topic decision-making in more depth. There are seven possible research projects that could continue this work.

The first research project could be to pilot the topic relevance questions in Table 48 with organisations developing tests. The questions could be fine-tuned and expanded based on discussions with the organisations. The final product could be an addendum of enhancements to the socio-cognitive framework.

The second research project could be to probe further into how on-topic background rating experience may affect on-topic decisions as discussed in the topic development section (5.1.1). The project would be two-fold: (1) raise awareness in

raters about different approaches to topic development as discussed in section 5.2.1.2 (Rater guidance and support materials); and (2) confirm whether raters are influenced by using different rating scales.

In this second project, a suitable method could be to cross-reference the raters' qualitative data in this thesis with the tests they have previously marked (I have collected this data) and then follow up with grounded theory interviews with raters on the types of topic development training they have received. This research project could also try to understand if teaching background influences their topic development perceptions as well as probe for affective schemata regarding topic development expectations for different task question prompts. This would be a type of member-checking research project recommended by Creswell and Plano Clark (2018).

The third research project would be similar to Burton's (2020) performance-evaluative and semantic differentials research using a 7-point Likert scale documented at the beginning of this Discussion chapter. After raters have marked selected spoken performances, they could choose how important different elements of topic relevance were for deciding on topic relevance for each test-taker response. Prior to collecting perceptions on each test-taker response, a benchmark questionnaire on raters' views about topic relevance should be collected. Follow-up interviews could investigate the reasons for the decisions, for example to help determine if rating scale experience, teaching experience, affective reactions to certain prompts, or something else are reasons for topic development decisions.

A fourth project could be to investigate how these findings could be useful for automated assessment. Although beyond the scope of this current study, which

focused on human rating perceptions, the concepts I looked at may have relevance for the development of automated scoring approaches. Test-taker responses widely vary making it difficult for an automated system to determine topic relevance. Raters who used short perceptions that directly addressed the question prompts using vocabulary and raters who focused on the grammar requirements in the question prompts were making similar decisions to what an automated scoring system could accomplish. Recent automated assessment models use words and phrases in example responses to create probabilities of the response being on-topic (Wang & Evanini, 2020). Example responses need to be collected for each question prompt and as a critical mass of test-taker responses are fed into the system the prompts could over time start to be marked automatically with a focus on vocabulary and correct grammar for the prompt. Vocabulary proved to be an important way that raters used to make their decisions and grammar could be seen as important for the task requirements.

A fifth research project could be to investigate whether speaking about feelings is a B1- or B2-level performance, if raters can reliably rate feelings and what type of rater training would help standardise these decisions.

A sixth research project could be to research raters' perceptions to the full speaking test, rather than separated into prompts. This research could investigate how raters manage the test-takers' movement from topic to topic and/or sub-topics using a similar semantic profiles approach as Maton (2013). Because the questions were separate, investigation into how raters interacted with the semantic profiles of the test-taker responses was not possible.

A seventh project would be to replicate the current study and ask raters to listen only once to the recording and use only introspective notes (as opposed to retrospective notes) to keep the data collection more standardised. This approach would match live marking more authentically where most raters do not listen more than once. Raters could tick a box if they would have listened more than once.

5.6 Conclusion

Speaking on-topic is a complex construct that is difficult to define and measure given the subjective nature of decision-making. This exploratory study provided evidence on the internal criteria that raters in this study use to make on-topic decisions. The consistency and construct relevance of these decisions were explored, as well as how these decisions affect test scoring. I compared and contrasted raters across the on-topic decisions, on-topic perceptions and Aptis scoring to understand relationships in the data in order to speculate on why raters were making decisions in particular ways and to identify thesis conclusions. I followed up with email exchanges that I used to triangulate the data.

The findings showed that the yes/no binary topic relevance decision in the rating scales is mostly working as expected with raters using their on-topic decisions to fine-tune their B2 level decisions, with some outliers. However, the wide variety of perceptions of what is considered on- and off-topic means that on-topic decisions affect scoring. Raters who report lenient on-topic perceptions, can have lenient on-topic decisions, which can lead to lenient Aptis scores (and vice versa). See sections 4.3.1.1 and 4.3.1.2. It is notable though, that there are raters who do not use their topic relevance decisions to decide on their Aptis scores (see section 4.3.1.3). These findings extend the research by Eckes (2012) who found that raters show distinct

patterns of attaching importance to certain scoring criteria and show criterion-related bias.

This research highlights the importance of defining the topic relevance construct and provides enhancements to the socio-cognitive model to help testing professionals develop their topic relevance construct. Raters need to have explicit topic relevance criteria for each test they mark and be trained on the topic relevance construct so that they make consistent topic relevance decisions. There is a need to raise language assessment literacy regarding the construct of topic relevance. This thesis provides ideas on how this training could be managed.

The use of email exchanges as a data collection method worked well in this context. Raters exchange emails with rater managers about marking issues so it was an authentic way to collect data. A grounded approach with a smaller group of raters may have deepened the data and is one area for follow-up research.

This research provided me with evidence to make changes (e.g., more abstractness in the photos) and provided a plan for enhanced rater training. Moreover, there is a documented research agenda. It must also be kept in mind that this research was exploratory so the topic relevance decisions were challenging and these findings should not be generalised to live testing.

The broader significance of this research is that it shows the extent of rater values around the topic relevance construct when raters are given a chance to think about ‘why’ and ‘how’ they make their on-topic decisions. What struck me as significant to the testing community is how raters have different approaches to collecting evidence, either by focusing on topic development, evaluating the extent test-takers adhere to the words and photos in the question prompts and/or by identifying on-

topic vocabulary. In sum, raters have different perceptions of what it means to speak on-topic and there is a wide range of thresholds in the decision-making.

References

Alderson, J. C. (1993). Judgments in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research*. Teachers of English to Speakers of Other Languages, Inc.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2020). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, Vol. 38(1)*, 6-40. https://doi.org/10.1177/0265532220927487

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1981). The construct validation of the FSI oral interview. *Language Learning, 31*(1), 67-86. https://doi.org/10.1111/j.1467-1770.1981.tb01373.x

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The Modern*

Language Journal, 70(4), 380-390. https://doi.org/10.1111/j.1540-4781.1986.tb05294.x

Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly, 9*(3), 225-248.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*(3), 279-293. https://doi.org/10.1080/0969594X.2010.526585

Bazeley, P. (2009). Integrating data analyses in mixed methods research. *Journal of mixed methods research, 3*(3), 203-207. https://doi.org/10.1177/1558689809334443

Bazeley, P. (2013). *Qualitative data analysis: Practical strategies*. SAGE.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2-9. https://doi.org/10.1111/j.1745-3992.2012.00238.x

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. SAGE.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3(2), 77101*.

British Council. (2015). *EAQUALS Core Inventory for General English, 2nd Edition*. British Council. Retrieved March 6, 2023 from https://www.teachingenglish.org.uk/sites/teacheng/files/pub-british-council-eaquals-core-inventoryv2.pdf

British Council. (2018). *British Council Aptis speaking and writing corpus.* British Council.

British Council. (2023a). *The Aptis test helps students and universities achieve success*. British Council. Retrieved March 6, 2023 from https://www.britishcouncil.org/exam/aptis-10-years/aptis-university-mexico

British Council. (2023b). *Research grants in language assessment*. British Council. Retrieved September 10, 2023 from https://www.britishcouncil.org/exam/aptis/research/grants-and-awards

British Council. (2023c). *Which Aptis ESOL is right for me?* British Council. Retrieved March 6, 2023 from https://www.britishcouncil.es/en/exam/aptis/variants

Brown, A. (2000). *An investigation of the rating process in the IELTS oral interview*. IELTS Australia. Retrieved March 26, 2023 from https://www.ielts.org/-/media/research-reports/ielts_rr_volume03_report3.ashx

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1-25.

Brown, A., Iwashita, N., & McNamara, T. F. (2005). An examination of rater orientations and test-taker performance on EAP speaking tasks. *TOEFL ETS monograph series, MS 29*.

Brown, G., & Yule, G. (1983). *Discourse Analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9780511805226

Bryman, A. (2007). Barriers to integrating quantitative and qualitative research. *Journal of mixed methods research, 1*(1), 8-8. https://doi.org/10.1177/1558689806290531

Burton, J. D. (2018). *Raters' perceptions and operationalization of authentic engagement in oral proficiency tests.* Master's thesis, University of Lancaster.

Burton, J. D. (2020). Raters' measurement of test-task authentic engagement in L2 oral-performance assessment: An exploration of scale development. *System, 90*, 102233. https://doi.org/10.1016/j.system.2020.102233

Bygate, M. (1987). *Speaking*. Oxford University Press.

Cai, H. (2015). Weight-Based Classification of Raters and Rater Cognition in an EFL Speaking Test. *Language Assessment Quarterly, 12*(3), 262-282. https://doi.org/10.1080/15434303.2015.1053134

Cambridge Dictionary. (2023). Retrieved January 14, 2023 from https://dictionary.cambridge.org/dictionary/english

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication*. Longman Group.

Canale, M. (1987). The measurement of communicative competence. *Annual Review of Applied Linguistics, 8*, 67-84. https://doi.org/10.1017/S0267190500001033

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47. https://doi.org/10.1093/applin/I.1.1

Caracelli, V., J., & Greene, J., C. (1993). Data analysis strategies for mixed-method evaluation designs. *Educational evaluation and policy analysis, 15*(2), 195-207. https://doi.org/10.2307/1164421

Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical development and integrated arguments*. Equinox.

Chapelle, C. (1999). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge University Press.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. M.I.T. Press.

Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. *Cambridge University Press*.

Council of Europe. (2018). Commmon European framework of reference for languages: Companion volume with new descriptors. *Council of Europe*.

Cox, T. L. (2017). Understanding intermediate-level speakers' strengths and weaknesses: An examination of OPI tests from Korean learners of English. *Foreign Language Annals, 50*(1), 84-113. https://doi.org/10.1111/flan.12258

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

Cumming, A., Kantor, R., & Powers, D. E. (2001). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67-96. https://doi.org/10.1111/1540-4781.00137

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135.

de Jong, N. (2023). Assessing second language speaking proficiency. *Annual* Review of Linguistics, 9, 541-560.

de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5-34. https://doi.org/10.1017/S0272263111000489

Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing, 32*(4), 521-541. https://doi.org/10.1177/0265532215575626

Dörnyei, Z. (2007). *Research methods in applied linguistics: quantitative, qualitative, and mixed methodologies*. Oxford University Press.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*(2), 125-144. https://doi.org/10.1177/026553229401100203

Dunn, K. (2024). Aptis scoring system version 2.1. *Aptis Technical Report TR/2024/002. British Council. London.* Retrieved 11 April, 2025 from https://www.britishcouncil.org/exam/english/aptis/research/publications/technical/aptis-scoring-system-v2-1

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A Many-Facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270-292.

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang GmbH.

Elder, C., Barkhuizen, G., Knoch, U., & Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64.

Elder, C., McNamara, T., Kim, H., Pill, J., & Sato, T. (2017). Interrogating the construct of communicative competence in language assessment contexts: What the non-language specialist can tell us. *Language and Communication, 57*. https://doi.org/10.1016/j.langcom.2016.12.005

ETS. (2023). *TOEFL iBT Independent Speaking Rubrics*. Retrieved 15 January 2023 from https://www.ets.org/content/dam/ets-org/pdfs/toefl/toefl-ibt-speaking-rubrics.pdf

Fairbairn, J. (2015a). *An examination of how analystic raters adapt to holistic marking*. Master's thesis, University of Lancaster.

Fairbairn, J. (2015b). *Maintaining marking consistency in a large-scale international test: The Aptis experience* [Poster presentation]. European Association for Language Testing and Assessment (EALTA) 12th Annual Conference, Copenhagen, Denmark.

Fairbairn, J., & Dunlea, J. (2017). Speaking and writing rating scales revision. *Aptis Technical Report TR/2017/001. British Council. London*.

Fairbairn, J., & Spiby, R. (2019). Towards a framework of inclusion: Developing accessibility in tests at the British Council. *European Journal of Special Needs Education, 34*(2), 236-255. https://doi.org/10.1080/08856257.2019.1581404

Fernandez, C. (2018). Behind a spoken performance: Test takers' strategic reactions in a simulated part 3 of the IELTS speaking test. *Language Testing in Asia, 8(18)*, 1-20. https://doi.org/10.1186/s40468-018-0073-4

Field, J. (2011). Cognitive validity. In L. B. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking*. Cambridge University Press.

Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing, 29*(3), 345-369. https://doi.org/10.1177/0265532211424479

Frost, K., Clothier, J., Huisman, A., & Wigglesworth, G. (2020). Responding to a TOEFL iBT integrated speaking task: Mapping task demands and test takers' use of stimulus content. *Language Testing*, 37(1), 133–155.

Frost, K., Wigglesworth, G., & Clothier, J. (2021). Relationships between Comprehension, Strategic Behaviours and Content-Related Aspects of Test Performances in Integrated Speaking Tasks. *Language Assessment Quarterly*, 18(2), 133-153.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208-238. https://doi.org/10.1177/026553229601300205

Fulcher, G. (2003). *Testing second language speaking*. Longman.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5-29. https://doi.org/10.1177/0265532209359514

Gates, G. E. (2018). *An analysis of rehearsed speech characteristics on the oral proficiency interview*. Masters thesis, Brigham Young University. Retrieved March 6, 2023 from https://scholarsarchive.byu.edu/etd/6731

Gigerenzer, G., & Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684-704.

Glaser, B. G. (1978). *Theoretical sensitivity: Advances in the methodology of grounded theory*. Sociology Press.

Glaser, B. G. (1992). *Basics of grounded theory analysis: Emergence vs forcing*. Sociology Press.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine Transaction.

Glaser, R. (1963). Instructional Technology and the Measurement of Learning Outcomes: Some Questions. *The American psychologist, 18*, 519. https://doi.org/10.1037/h0049294

Godfroid, A., & Hopp, H. (2023). *The Routledge handbook of second language acquisition and psycholinguistics*. Routledge. https://doi.org/10.4324/9781003018872

Goulden, N. R. (1992). Theory and vocabulary for communication assessments. *Communication Education, 41*(3), 258-269. https://doi.org/10.1080/03634529209378887

Green, R. (2013). *Statistical analyses for language testers*. Palgrave Macmillan.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational evaluation and policy analysis, 11*(3), 255-274. https://doi.org/10.3102/01623737011003255

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 29*(4), 759-762.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing, 3*(1), 49-68. https://doi.org/10.1016/1060-3743(94)90005-1

Hawkins, J. E. (2018). The practical utility and suitability of email interviews in qualitative research. *Qualitative report, 23*(2), 493.

Higgs, T., & Clifford, R. (1982). The push towards communication. In T. Higgs (Ed.), *Curriculum, Competence and the Foreign Language Teacher*. National Textbook Company.

Huang, H.-T. D., Hung, S.-T. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing, 35*(1), 27-49. https://doi.org/10.1177/0265532216677106

Huot, B. (1990a). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237-263. https://doi.org/10.3102/00346543060002237

Huot, B. (1990b). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(2), 201-213. https://doi.org/10.2307/358160

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Penguin Books.

IELTS. (2023). *Speaking: Band descriptors (public version)*. Retrieved January 15, 2023 from https://www.ielts.org/-/media/pdfs/ielts-speaking-band-descriptors.ashx.

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing, 33(3),* 341-366.

Inoue, C., & Lam, D. M. K. (2021). The effects of extended planning time on candidates' performance, processes, and strategy use in the lecture listening-into-speaking tasks of the TOEFL iBT® test. ETS Research Report No. RR-21-09.

Isaacs, T. (2016). Assessing Speaking. In D. Tsagair & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 131-146). DeGruyter Mouton.

Isaacs, T., & Thomson, R. I. (2020). Reactions to second language speech: Influences of discrete speech characteristics, rater experience, and speaker first language background. *Journal of Second Language Pronunciation, 6(3)*, 402-429.

Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied psycholinguistics, 32*(1), 113-140. https://doi.org/10.1017/S0142716410000317

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility. *Studies in Second Language Acquisition, 34*(3), 475-505. https://doi.org/10.1017/S0272263112000150

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics,* 29(1), 24-49. https://doi.org/10.1093/applin/amm017

Iwashita, N., May, L. & Moore, P. (2017). Features of discourse and lexical richness at different performance levels in the Aptis speaking test. *Assessment Research Awards and Grants Reports AR-G/2017/2. British Council. London*.

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes, 18*(3), 213-241. https://doi.org/10.1016/S0889-4906(97)00053-7

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Kahng, J. (2023). Exploring individual differences in rating second language speech: rater's language aptitude, marjor, accent familiarity, and attitudes. TESOL Quarterly, 57(4), 1545-1557.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kang, O. & Johnson, D. (2021). Linguistic features and automatic scoring of Aptis speaking performances. *Assessment Research Awards and Grants Reports AR-G/2021/3.* British Council. London.

Khabbazbashi, N. (2021). *On topic validity in speaking tests*. Cambridge University Press.

Khalifa, H., & Salamoura, A. (2011). Criterion-related validity. In L. B. Taylor (Ed.), *Examining speaking: Research and practice in assessing second langauge speaking* (Vol. 30, pp. 259-292). Cambridge University Press.

Kim, Y.-H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Language Testing, 28*(4), 509-541. https://doi.org/10.1177/0265532211400860

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior: A longitudinal study. *Language Testing, 28*(2), 179-200. https://doi.org/10.1177/0265532210384252

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35*(4), 477-499. https://doi.org/10.1177/0265532217710049

Knoch, U., Fairbairn, J., & Huisman, A. (2015). *An evaluation of the effectiveness of training Aptis raters online*. British Council. Retrieved on March 6, 2023 from https://www.britishcouncil.org/sites/default/files/knoch_fairbairn_and_huisman_0.pdf

Knoch, U., Fairbairn, J., & Huisman, A. (2016). *An evaluation of an online rater training program for the speaking and writing sub-tests of the Aptis test*. ALTAANZ. Retrieved on March 6, 2023 from https://www.altaanz.org/uploads/5/9/0/8/5908292/6._plta_5_1__knoch_et_al.pdf

Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: Issues, options and directions*. Equinox.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26-43.

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language*

Assessment Quarterly, 13(4), 377-392. https://doi.org/10.1080/15434303.2016.1237516

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing, 31(3), 329-348*.

Kuiken, F., & Vedder, I. (2016). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing, 34*(3), 321-336. https://doi.org/10.1177/0265532216663991

Kvale, S. (2007). *Doing interviews*. SAGE.

Kvale, S., & Svend, B. (2009). Conducting an interview. In S. Kvale & S. Brinkmann (Eds.), *Interviews: Learning the craft of qualitative research interviewing* (2nd ed ed.). SAGE.

Laborda, J.G., Pizarro, M.A., Litzler, M.F., Esteban, S.G. & Otero de Juan, N. (2017). Student perceptions of the CEFR levels and the impact of guided practice on Aptis oral test performance. *Assessment Research Awards and Grants Reports AR-G/2017/5.* British Council. London.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing, 13*(2), 151-172. https://doi.org/10.1177/026553229601300202

Lee, Y.-W. (2005) Dependability of scores for a new ESL speaking test: evaluating prototype tasks. *TOEFL Monograph MS-28.* ETS.

Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*(2), 131-166. https://doi.org/10.1191/0265532206lt325oa

Leung, C. (2010). English as an additional language: learning and participating in manistream classrooms (pp. 182-205). In P. Seedouse et al. (Ed.), *Conceptualising 'learning' in applied linguistics*. Palgrave Macmillan.

Levelt, W. J. M. (1989). *Speaking: from intention to articulation*. MIT Press.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543-560.

Linacre, J. M. (1989). *Many-Facet Rasch Measurement*. Mesa Press.

Linacre, J.M. (2002). Opimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement,* 3(1), 85-106. Retrieved 23 October 2024 from https://www.researchgate.net/profile/John-Linacre/publication/11372384_Understanding_Rasch_measurement_Optimizing_rating_scale_category_effectiveness/links/65548bccb1398a779d8f59be/Understanding-Rasch-measurement-Optimizing-rating-scale-category-effectiveness.pdf

Linacre, J. M. (2018a). *Many-Facets Rasch measurement software, v*ersion 4.2.1. available online: www.winsteps.com.

Linacre, J. M. (2018b). *A user's guide to Facets Rasch-model computer programs.* In www.winsteps.com.

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31*(4), 479-499. https://doi.org/10.1177/0265532214530699

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

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking.

Language Testing, 22(4), 415-437. https://doi.org/10.1191/0265532205lt303oa

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15(2)*, 158-180.

Mackey, A., & Gass, S. M. (2015). *Second language research: Methodology and design*. Taylor & Francis Group.

McKinley, J., & Rose, H. (2020). *The Routledge handbook of research methods in applied linguistics*. Routledge.

McNamara, T. F. (1996). *Measuring second language performance*. Longman.

McNamara, T. F., Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics, 22*, 221-242. https://doi.org/10.1017/S0267190502000120

McNamara, T.F., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment.* Oxford University Press.

Meho, L. (2006). E-mail interviewing in qualitative research: A methodological discussion. *Journal of the American Society for Information Science and Technology, 57*(10), 1284-1295. https://doi.org/10.1002/asi.20416

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13-103). American Council on Education and Macmillan.

Messick, S. (1994). The matter of style: Manifestations of personality in cognition, learning, and teaching. *Educational Psychologist, 29*(3), 121-136. https://doi.org/10.1207/s15326985ep2903_2

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Ed.), *Performance testing, cognition, and assessment: selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem* (pp. 92-114). Cambridge University Press.

Myford, C., & Wolfe, E. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of Applied Measurement, 4(4), 386-422.*

Myford, C., & Wolfe, E. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5(2), 189-227*.

Nakatsuhara, F. (2008). Inter-interviewer variation in oral interview tests. *ELT Journal, 62*(3), 266-275.

Norris, J. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics, 36*, 230-244. https://doi.org/10.1017/S0267190516000027

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19*(4), 395-418. https://doi.org/10.1191/0265532202lt237oa

North, B. (2003). *Scales for rating languages performance: Descriptive models, formulation styles, and presentation formats*. Education Testing Service.

O'Hagan, S & Frost, K. (2020). Examining test-taker processes and strategies and stakeholder perceptions of relevance of the Aptis for Teachers speaking test in the Australian context. *Assessment Research Awards and Grants Reports AR-G/2020/1.* British Council. London.

Orr, M. (2002). The FCE Speaking test: using rater reports to help interpret test scores. System, 30, 143-154.

O'Sullivan, B. (2012). Assessing Speaking. In C. Coombe (Ed.), *The Cambridge guide to second language assessment*. Cambridge University Press.

O'Sullivan, B. (2015). Aptis test development approach. British Council. Retrieved 29 March, 2024 from https://www.britishcouncil.org/exam/english/aptis/research/publications/test-development-approach

O'Sullivan, B., Dunlea, J., Spiby, R., Westbrook, C., & Dunn, K. (2020). *Aptis General technical manual version 2.2*. British Council. Retrieved on 6 March, 2023 from https://www.britishcouncil.org/exam/aptis/research/publications/technical/general-technical-manual-version-2-2

O'Sullivan, B., & Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: Theory and practice (pp. 13-33)*. Palgrave.

O'Sullivan, B., & Chalhoub-Deville, M. (2021). Language testing for migrants: Co-constructing validation. *Language Assessment Quarterly, 18(5)*, 1-11. https://doi.org/10.1080/15434303.2021.1986513

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590-601. https://doi.org/10.1093/applin/amp045

Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly, 36*(2), 219-233. https://doi.org/10.2307/3588333

Patton, M. Q. (1980). *Qualitative evaluation and research methods* (2nd ed.). SAGE.

Penfield, R. D. (2014). An NCME Instructional Module on Polytomous Item Response Theory Models. *Educational measurement, issues and practice, 33*(1), 36-48. https://doi.org/10.1111/emip.12023

Polkinghorne, D. E. (2005). Language and meaning: Data collection in qualitative research. *Journal of Counseling Psychology, 52*(2), 137-145. https://doi.org/10.1037/0022-0167.52.2.137

Qiu, X., & Bui, G. (2022). Pre-task planning effects on learner engagement in face-to-face and synchronous computer-mediated communication. *Language Teaching Research*. https://doi.org/10.1177/13621688221135280

Ratislavová, K., & Ratislav, J. (2014). Asynchronous email interview as a qualitative research method in the humanities. *Humaff, 24*(4), 452-460. https://doi.org/10.2478/s13374-014-0240-y

Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics, 37*(6), 828-848. https://doi.org/10.1093/applin/amu069

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: a triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction*. Cambridge University Press.

Robinson, P. (2015). The cognitive hypothesis, second languge task demands, and the SSARC model of pedagogic task sequencing. In M. Bygate (Ed.), *Domains and Directions in the Development of TBLT. A decade of plenaries from the international conference*. John Benjamins Publishing Company.

Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing, 35*(3), 331-355.

Rose, H., McKinley, J., & Baffoe-Djan, J.B. (2019). Introspective and Retrospective Protocols. *In Data Collection Research Methods in Applied Linguistics* (pp.

49–66). Bloomsbury Academic. Retrieved March 25, 2024, from http://dx.doi.org/10.5040/9781350025875.ch-003

Rossi, O. (2021). *Pre-testing Aptis for teachers speaking/writing test versions (unpublished report)*. British Council.

Spicer, D. P., & Sadler-Smith, E. (2005). An examination of the general decision making style questionnaire in two UK samples. https://doi.org/10.1108/02683940510579777

Saldana, J. (2016). *The coding manual for qualitative researchers* (3rd ed.). SAGE.

Sasayama, S. (2016). Is a 'complex' task really complex? Validating the assumption of cognitive task complexity. *The Modern Language Journal*, 100(1), 231–254.

Sasayama, S., & Norris, J. M. (2022). Designing speaking tasks for different assessment goals: The complex relationship between cognitive task complexity, language performance, and task accomplishment. *TASK: Journal on Task-Based Language Teaching and Learning, 2*(2), 184-217. https://doi.org/10.1075/task.21020.sas

Seedhouse, P., & Harris, A. (2011). Topic development in the IELTS Speaking Test. *IELTS Research Reports, 12*. Retrieved on March 6, 2023 from https://www.ielts.org/-/media/research-reports/ielts_rr_volume12_report2.ashx

Shaw, S. (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. *Cambridge Research Notes, 8 (5), 13-17*.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge University Press.

Sheehan, S., Sanderson, P. & Harris, A. (2017). Identifying key criteria of written and spoken English at C1. *Assessment Research Awards and Grants Reports AR-G/2017/1.* British Council. London.

Shin, J., Rodríguez-Fuentes, R. A., Swatek, A. M., & Ginther, A. (2022). *Aptis test review.* Language Testing, 39(1), 172–187. https://doi.org/10.1177/02655322211032873

Shohamy, E. (1983). Rater reliability of the oral interview speaking test. *Foreign Language Annals, 16*(3), 219-222. https://doi.org/10.1111/j.1944-9720.1983.tb01456.x

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Spicer, D. P., & Sadler-Smith, E. (2005). An examination of the general decision making style questionnaire in two UK samples. *Journal of Managerial Psychology. 20, 1/2.* ProQuest https://doi.org/10.1108/02683940510579777

Spolsky, B. (1990). Oral examinations: An historical note. *Language Testing, 7*(2), 158-173. https://doi.org/10.1177/026553229000700203

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford University Press.

Stansfield, C. W., & Kenyon, D. M. (1992a). The development and validation of a simulated oral proficiency interview. *The Modern Language Journal, 76*(2), 129-141. https://doi.org/10.1111/j.1540-4781.1992.tb01093.x

Stansfield, C. W., & Kenyon, D. M. (1992b). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20*(3), 347-364. https://doi.org/10.1016/0346-251X(92)90045-5

Sweedler-Brown, C. O. (1985). The influence of training and experience on holistic essay evaluations. *The English Journal, 74*(5), 49-55. https://doi.org/10.2307/817702

Tavakoli, P, Nakatsuhara, F. & Hunter, A. (2017). Scoring validity of the Aptis speaking test: investigating fluency across tasks and levels of proficiency.

Assessment Research Awards and Grants Reports AR-G/2017/7. British Council. London.

Taylor, L., & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (Vol. 30). Cambridge University Press.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. SAGE.

Thunholm, P. (2004). Decision-making style: habit, style or both? *Personality and Individual Differences, 36*(4), 931-944. https://doi.org/10.1016/S0191-8869(03)00162-4

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3-12. https://doi.org/10.1093/elt/49.1.3

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing, 16*(1), 82-111. https://doi.org/10.1177/026553229901600105

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 325-344. https://doi.org/10.1177/0265532211424478

Wang, X., & Evanini, K. (2020). Features measuring content and discourse coherence. In *Automated Speaking Assessment* (1 ed., pp. 138-156). Routledge. https://doi.org/10.4324/9781315165103-9

Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly, 12*(3), 283-304. https://doi.org/10.1080/15434303.2015.1037446

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

Weigle, S. C. (1998). Using Facets to model rater training effects. *Language Testing, 15*(2), 263-287.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Weir, C.J., Vidakovic, I., & Galaczi, E. (2013). *Measured constructs: a history of Cambridge English language examinations 1913-2012*. Cambridge University Press.

Widdowson, H. G. (1989). Knowledge of language and ability for use. *Applied Linguistics, 10*(2), 128-137. https://doi.org/10.1093/applin/10.2.128

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*(3), 305-319.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 85-106. https://doi.org/10.1177/026553229701400105

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing, 35*(2), 161-192. https://doi.org/10.1177/0265532216686999

Wolcott, H. F. (1994). *Transforming qualitative data: Description, analysis, and interpretation*. SAGE.

Wolfe, E. W. (2006). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment, Vol. 2, No. 1, pp. 37-56*.

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substatively interesting raters. *Educational Measurement: Issues and Practice*, 31(3), 31-37.

Wright, B. D., and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 8, 370. Retrieved 8 March 2025 from https://www.rasch.org/rmt/rmt83b.htm.

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning, 61*(4), 1222-1255. https://doi.org/10.1111/j.1467-9922.2011.00667.x

Yan, X., Kim, H.R. & Kim, J.Y. (2018). Complexity, accuracy and fluency (CAF) features of speaking performances on Aptis across different levels on the common European framework of reference (CEFR) report. *Assessment Research Awards and Grants Reports AR-G/2018/1.* British Council. London.

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31-50. https://doi.org/10.1177/0265532209360671

Appendix 1: Aptis General Speaking test

Part	Skill focus	Lvl	Task description	Channel of input / prompts	Time to plan	Time for response	Rating criteria
1	**Giving personal information**	**A1/A2**	Candidate responds to 3 questions on personal topics. The candidate records his/her response before the next question is presented.	Questions presented in both written and oral form (pre-recorded). Questions presented in a sequence (e.g. Q2 is presented after the response to Q1).	No	30 seconds to respond to each question	Separate task-based holistic scales are used for each task. Performance descriptors describe the expected performance at each score band. The following aspects of performance are addressed: *1) grammatical range and accuracy* *2) lexical range and accuracy* *3) pronunciation* *4) fluency* *5) cohesion and coherence.*
2	**Describing, expressing opinions, providing reasons and explanations**	**B1**	The candidate responds to 3 questions. The first asks the candidate to describe a photograph. The next two are on a concrete and familiar topic related to the photo.	1) Questions presented in both written and oral form (pre-recorded). Questions presented in a sequence (e.g. Q2 is presented after the response to Q1). 2) A single photo of a scene related to the topic and familiar to A2/B1 candidates on screen.	No	45 seconds to respond to each question	
3	**Describing, comparing and contrasting, providing reasons and explanations**	**B1**	The candidate responds to 3 questions / prompts and is asked to describe, contrast and compare two photographs on a topic familiar to B1 candidates. The candidate gives opinions and provides reasons and explanations.	1) Questions presented in both written and oral form (pre-recorded). Questions presented in a sequence (e.g. Q2 is presented after the response to Q1). 2) Two photographs showing different aspects of a topic are presented on screen.	No	45 seconds to respond to each question	
4	**Integrating ideas on an abstract topic into a long turn. Giving and justifying opinions, advantages and disadvantages**	**B2**	The candidate plans a longer turn integrating responses to a set of 3 questions related to a more abstract topic. After planning their response, the candidate speaks for two minutes to present a coherent, continuous, long turn.	1) Three questions are presented simultaneously in both written and oral form (pre-recorded). Questions remain on screen throughout the task. 2) One photograph illustrating an element of the topic mentioned in the prompts. The photo is not referred to in the questions.	1 minute	2 minutes for the entire response, integrating the 3 questions into a single long turn	

Appendix 2: Aptis Speaking B2 task versions used in this research

These question prompts are available to researchers.

Version 1

Part Four - In this part I am going to show you a picture and ask you three questions. You will have one minute to think about your answers before you start speaking. You will have two minutes to answer all three questions. Begin speaking when you hear this sound (beep). Look at the photograph.

Tell me about an outdoor activity you did as a child.

How did you feel about doing this activity?

Is it important for children to do outdoor activities?

You now have one minute to think about your answers. You can make notes if you wish.

You now have two minutes to talk.

Version 2

Part Four - In this part I am going to show you a picture and ask you three questions. You will have one minute to think about your answers before you start speaking. You will have two minutes to answer all three questions. Begin speaking when you hear this sound (beep). Look at the photograph.

Tell me about a time when you went on a long journey.

How did you feel about it?

How do people travel long distances in your country?

You now have one minute to think about your answers. You can make notes if you wish.

You now have two minutes to talk.

Version 3

Part Four - In this part I am going to show you a picture and ask you three questions. You will have one minute to think about your answers before you start speaking. You will have two minutes to answer all three questions. Begin speaking when you hear this sound. Look at the photograph.

Tell me about a time when you visited a very tall building.

How did you feel about it?

Why do you think so many cities have tall buildings?

You now have one minute to think about your answers. You can make notes if you wish.

You now have two minutes to talk.

Appendix 3: Aptis Speaking Task 4 specifications

Test	Aptis	Component	Speaking	Task	Task 4

Features of the Task						
Skill focus	Integrating ideas regarding an abstract topic into a long turn. Giving opinions, justifying opinions, giving advantages and disadvantages.					
Task level (CEFR)	A1	A2	B1	B2	C1	C2
Task description	The candidate plans a long turn integrating responses to a set of 3 questions related to a more abstract topic. The candidate speaks for two minutes to present his/her long-turn. The 3 questions expand in focus and cognitive demand (see features of the input/prompts below).					
Task description: extra information	The task requires a long turn response in relation to abstract topics. The illustration is only for additional contextualisation of the topic and is not referred to directly in any of the questions.					
Instructions to candidates	*Part four. In this part, I'm going to show you a picture and ask you three questions. You will have one minute to think about your answers before you start speaking. You will have two minutes to answer all three questions. Begin speaking when you hear this sound (beep). Look at the photograph.*					
Presentation of rubric	Aural		Written		Visual non-verbal (e.g. photo)	
Response format	Q&A		Short turn		Long turn	
Planning time	1 minute					
Delivery	Face-to-face	Telephone	Computer	Other		
Nature of input	Real time (face-to-face)	Real time (remote)	Pre-recorded input	No aural input		
	Unscripted	Guided	Semi-scripted	Scripted	N/A	
Nature of interaction	Interlocutor–Candidate (I–C)		Candidate–Candidate (C–C)			
	Candidate only (C)		Interlocutor–Candidate–Candidate			
Functions targeted	**Informational Functions**		**Interactional Functions**		**Managing Interaction**	
	Providing personal information		Agreeing			
	Explaining opinions/preferences		Disagreeing		Initiating	
	Elaborating		Modifying/ commenting		Changing topics	
	Justifying opinions		Asking for opinions		Reciprocating	
	Comparing		Persuading		Deciding	
	Speculating		Asking for information			
	Staging		Conversational repair			
	Describing		Negotiation of meaning			
	Summarising					
	Suggesting					
	Expressing preferences					

Features of the Input / Prompt										
Description	Three questions. 1) Asks for a description of personal experience in relation to an abstract topic; 2) Asks for elaboration on the candidate's impression/opinion in relation to the topic; 3) Asks for a more objective discussion of the topic from the perspective of wider relevance to society/people in general. A photograph is provided for extra contextualisation but is not referred to in the questions.									
Length of questions	Maximum of 20 words per question									
Lexical level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Grammatical level	A1–B1 Grammatical exponents (See Guidelines on Adhering to Grammatical Level).									
Content knowledge	General									Specific
Cultural specificity	Neutral									Specific
Nature of information	Only concrete		Mostly concrete		Fairly abstract		Mainly abstract			
Relevant domain	Public		Occupational		Educational		Personal			
Topic	From topic list for B2.									

Features of the Expected Response			
Description	A long turn of 2 minutes. Candidate must provide a coherent and cohesive long turn which deals with at least 2 questions to achieve a rating of 3 (out of 5) for the task.		
Length of response	Up to 2 minutes for the entire long turn. Adequate length for B2-level performance will generally require the candidate to speak for the full two minutes or most of the full two minutes.		
Lexis /grammar	Demonstration of grammatical control at the B2 level necessary for a rating of 3 (out of 5) for the task. B2 lexis sufficient to respond adequately to all questions.		
Rating scale for task	A task-specific holistic rating scale is used for the task. The rating scale is a 6-point scale from 0–5. A B2-level performance is required to achieve score bands 3–4. A score of 5 is awarded for performances beyond B2 level.		
Timing of rating	Real time		After test event
Rater	Interlocutor	Rater present at test	Rater not present at test event

Appendix 4: Speaking B2 rating scale

6 **C2**	Likely to be above C1 level.
5 **C1**	Response addresses all **three** questions and is well-structured. • Uses a range of complex grammar constructions accurately. Some minor errors occur but do not impede understanding. • Uses a range of vocabulary to discuss the topics required by the task. Some awkward usage or slightly inappropriate lexical choices. • Pronunciation is clearly intelligible. • Backtracking and reformulations do not fully interrupt the flow of speech. • A range of cohesive devices are used to clearly indicate the links between ideas.
4 **B2.2**	Responses to all **three** questions are on-topic and show the following features • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. • Pronunciation is intelligible. Mispronunciations do not put a strain on the listener or lead to misunderstanding • Some pausing while searching for vocabulary but this does not put a strain on the listener. • A limited number of cohesive devices are used to indicate the links between ideas.
3 **B2.1**	Responses to **two** questions are on-topic and show the following features • Some complex grammar constructions used accurately. Errors do not lead to misunderstanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding. • Pronunciation is intelligible. Mispronunciations do not put a strain on the listener or lead to misunderstanding • Some pausing while searching for vocabulary but this does not put a strain on the listener. • A limited number of cohesive devices are used to indicate the links between ideas.
2 **B1.2**	Responses to at least **two** questions are on-topic and show the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Limitations in vocabulary make it difficult to deal fully with the task. • Pronunciation is intelligible but occasional mispronunciations put an occasional strain on the listener. • Noticeable pausing, false starts, reformulations and repetition. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
1 **B1.1**	Response to **one** question is on-topic and shows the following features • Control of simple grammatical structures. Errors occur when attempting complex structures. • Limitations in vocabulary make it difficult to deal fully with the task. • Pronunciation is intelligible but occasional mispronunciations put an occasional strain on the listener. • Noticeable pausing, false starts, reformulations and repetition. • Uses only simple cohesive devices. Links between ideas are not always clearly indicated.
0 **A1/A2**	Performance not sufficient for B1, or no meaningful language, or the responses are completely off-topic (memorised or guessing).

Appendix 5: UCL ethics approval

Institute of Education

UCL

Doctoral Student Ethics Application Form

Anyone conducting research under the auspices of the Institute (staff, students or visitors) where the research involves human participants or the use of data collected from human participants, is required to gain ethical approval before starting. This includes preliminary and pilot studies. Please answer all relevant questions in simple terms that can be understood by a lay person and note that your form may be returned if incomplete.

***Registering your study with the UCL Data Protection Officer as part of the UCL Research Ethics Review Process**

If you are proposing to collect personal data i.e. data from which a living individual can be identified **you must be registered with the UCL Data Protection Office before you submit your ethics application for review**.

If the Data Protection Office advises you to make changes to the way in which you propose to collect and store the data this should be reflected in your ethics application form.

For further information see Steps 1 and 2 of our Procedures page at:
https://ethics.grad.ucl.ac.uk/procedures.php

Section 1 Project details					
a.	Project title			Communicative adequacy in the Aptis test at B2 level	
b.	Student name and ID number (e.g. ABC12345678)			Judith Fairbairn 17056231	
c.	*UCL Data Protection Registration Number			Z6364106/2018/03/128 26 March 2018	
c.	Supervisor/Personal Tutor			Talia Isaacs	
d.	Department			Culture, Communication & Media	
e.	Course category (Tick one)	PhD	☒	EdD	☐
		DEdPsy	☐		
f.	**If applicable,** state who the funder is and if funding has been confirmed.			British Council part funding	
g.	Intended research start date			4 April 2017	
h.	Intended research end date			3 April 2022	
i.	Country fieldwork will be conducted in *If research to be conducted abroad please check www.fco.gov.uk and submit a completed travel risk assessment form (see guidelines). If the FCO advice is against travel this will be **required** before ethical approval can be granted: http://ioe-net.inst.ioe.ac.uk/about/profservices/international/Pages/default.aspx*			UK	
j.	Has this project been considered by another (external) Research Ethics Committee?				
	Yes ☐	External Committee Name:			
	No ☒ ⇨ *go to Section 2*	Date of Approval:			

Appendix 6: Email to raters

Dear Examiner,

This research project is part of the ongoing Aptis test validation work to ensure the Aptis test, the rating scales and Aptis rater training materials are continuously reviewed and improved.

Aptis raters who join this project will review 45 Aptis General Task 4 responses, fill in an Excel spreadsheet with their written perceptions of the responses, and attend an email interview to discuss their perceptions in more detail. The amount of time you would need to devote to this project is 5 hours and you would complete the work from home. You will complete the work in at least three blocks of time of your choosing within deadline.

You will have two weeks to complete the perceptions (start date can be any time from July – November), which can be done in one sitting (approximately 3 hours). You will then engage in an email exchange interview, which will take place in the two weeks following receipt of your Excel spreadsheet. There will be a series of emails (at least two) and you must be able to respond to the emails within seven days. Each email will take approximately 30 minutes to one hour to answer.

In the event that there are more than 35 Aptis raters who would like to participate, raters will be ranked by length of service (in years and months) and those with more service will be selected.

Raters will be paid the senior rater hourly rate for four hours of work and payment will be in the month when the work is completed.

Research will take place from July to November and start dates are flexible so please email me with an expression of interest if you would like to participate.

Appendix 7: Participant information sheet

Participant Information Sheet For Aptis Examiners

UCL Research Ethics Committee Approval ID Number: Z6364106/2018/03/128 (26 March 2018)

Title of Study: Communicative competence in the Aptis test

Department: Institute of Education, Department of Culture, Communication & Media

Name and Contact Details of the Researcher: Judith Fairbairn

Name and Contact Details of the Research Supervisor: Dr. Talia Isaacs

1. Invitation to participate in a PhD research project on the Aptis test

You are being invited to take part in a PhD research project on the communicative competence of language learners. This project will look at the quality of responses to Aptis General Speaking Task 4 (B2 task). Before you decide to participate, it is important for you to understand why the research is being done and what participation will involve. Please take time to read the following information carefully and discuss it with me, Judith Fairbairn,the Examiner Network Manager, Catherine Hughes, the Head of Assessment Services, Jonathan Ford, or my PhD supervisor, Talia Isaacs if you wish. Ask us if there is anything that is unclear or if you would like more information. Take time to decide whether or not you wish to take part. Thank you for reading this document.

2. What is the project's purpose?

This research project is part of the ongoing Aptis test validation work to ensure the Aptis test, the rating scales and examiner training materials are continuously reviewed and improved. This research may also contribute to the theoretical model of communicative competence in second language acquisition.

Aptis examiners who join this project will mark 30 Aptis General Task 4s (B2 task), write reasons (explanations) for their marks in an Excel spreadsheet and attend an interview to discuss the marks in more detail. The amount of time you would need to devote to this project is 3 hours and you would complete the work from home. You can complete the work in one sitting or in blocks of time of your choosing within deadline.

The work will take place in two parts:

- Part 1: Complete an Excel spreadsheet with decisions and explanations for decisions – 3 hours

- Part 2: Email exchange interviews – 2 hours

Examiners will be paid the Senior Examiner hourly rate for five hours of work.

3. Why have I been chosen?

As an active Aptis Examiner you have been chosen to apply to work on this project. Twenty-five Aptis Examiners will take part in this research study and all Examiners with at least one year of rating experience and normally marking for at least five hours per month have been invited to participate. In the event that there are more than 35 raters who would like to participate, Examiners will be ranked by length of service (in years and months) and those with more service will be selected.

4. Do I have to take part?

It is up to you to decide whether or not to take part and participation is entirely voluntary. If you decide to take part you will be asked to sign a consent form. You can withdraw from the research at any time without giving a reason.

If you decide not to participate in Part 2 (June interview), you will still be paid for your work in Part 1 (decisions and explanations). If you decide to withdraw before the interview, you will be asked what you wish to happen to your decisions and

explanations with the option to delete the data. You can also withdraw your interview data up to 4 weeks after the interview and still be paid for the interview.

5. What will happen to me if I take part?

1. The Researcher will send you a consent form to print, sign, scan and return by email.

2. You will next fill in a short questionnaire by Confirmit, the British Council's secure questionnaire system in order to collect information about your testing and teaching background, other languages you know and your Aptis task marking preferences.

3. The Researcher will next send you the 45 speaking Task 4 responses to review. You will fill in an Excel spreadsheet with your decisions and comments. These will be sent in a secure file transfer system. Once you have completed the marking, the spreadsheet will be sent to the Researcher using the secure file transfer system. You are paid for three hours to complete the review. You can complete the marking at any time during the month of _____(month to be agreed).

4. The Researcher will contact you by email to arrange an email exchange interview. The email exchange interview will take two hours and be a series of up to three emails. Some of the responses will be selected for further comment by you. There are no correct or incorrect answers to the questions and you can change your mind after listening again. The Researcher is genuinely interested in your views on the test-taker performance.

You may be invited for a follow-up interview, focus group or email exchange. You can choose to join future research (with payment) if you wish. Participation in future research is entirely voluntary.

5. Will I be recorded and how will the recorded media be used?

The audio recordings from any follow-up video conference interviews will be recorded and transcribed. The transcriptions will be anonymous and used only for analysis and for illustration in the PhD dissertation or conference presentations. Your voice recordings will never be played in public. No other use will be made of them without your written permission, and no one outside the project will be allowed access to the original recordings. Once the files have been transcribed and the PhD completed, the recordings will be deleted.

6. What are the possible disadvantages and risks of taking part?

There shouldn't be any disadvantages from taking part in this research and data will not be used to judge your performance as a rater. If any unexpected disadvantages arise during the research, you can bring it immediately to the attention of the Researcher (Judith Fairbairn), the UCL Research Supervisor (Talia Isaacs) or the Aptis Examiner Network Manager (Catherine Hughes).

7. What are the possible benefits of taking part?

You will be paid for three hours of your time and be part of a team contributing to the validation of the Aptis test. Your participation will help us improve the Aptis test, the rating scales and training materials.

It is up to you to decide whether or not to take part and participation is entirely voluntary. If you decide to take part you will be asked to sign a consent form. You can withdraw from the research at any time without giving a reason.

If you decide not to participate in Part 2 (email exchange interview), you will still be paid for your work in Part 1 (review and decisions). If you decide to withdraw before the interview, you will be asked what you wish to happen to your decisions and explanations with the option to delete the data. You can also withdraw your interview data up to 4 weeks after the interview and still be paid for the interview.

8. What if something goes wrong?

Entering data into the Excel spreadsheet template, the file transfer system and the on-line interview software (if invited for a video conference interview) are areas that could possibly go wrong (for example forgetting to save the document or internet connectivity issues). However, Aptis examiners work remotely and online as part of their Aptis examining work so you shouldn't encounter any problems. The Researcher will be on email to provide any technical support.

If something does go wrong and you feel your complaint has not been handled to your satisfaction by the Researcher, you can contact the UCL Supervisor or the Chair of the UCL Research Ethics Committee.

9. Will my taking part in this project be kept confidential?

All the information that we collect about you during the course of the research will be kept strictly confidential and will not be shared with Examiner management staff. You will not be able to be identified in any ensuing reports or publications. All transcriptions will be done by the Researcher (not outsourced) and anonymised.

10. Limits to confidentiality

There are no limits to confidentiality. All data will be confidential only to the Researcher and quotes from the explanations or interviews will be used anonymously in reports.

11. What will happen to the results of the research project?

The results will be written up in a PhD thesis and the Researcher is more than happy to share the results with participants.

Once the PhD has been accepted, the interview voice recordings will be deleted.

12. Data Protection Privacy Notice

The data controller for this project is the British Council. The British Council Information Governance & Risk Management office provides oversight of British Council activities involving the processing of personal data.

Your personal data will be processed for the purposes outlined in this notice. The legal basis that would be used to process your personal data will be the provision of your consent. You can provide your consent for the use of your personal data in this project by completing the consent form that has been provided to you.

Your personal data will be processed so long as it is required for the research project (expected completion of PhD is 2022). Your interview data will be deleted once the PhD has been completed. Your decisions and explanations and transcript of the interview will be anonymised to minimise the processing of personal information. No data will be transferred outside the EEA.

If you are concerned about how your personal data is being processed, please contact UCL in the first instance. If you remain unsatisfied, you may wish to contact the Information Commissioner's Office (ICO). Contact details, and details of data subject rights, are available on the ICO website.

13. Who is organising and funding the research?

The British Council and the Researcher are jointly funding this research.

16. Contact for further information

For further information please contact the Researcher or the Research Supervisor:

Judith Fairbairn
PhD Researcher and Assessment Development Manager at the British Council
UCL Centre for Applied Linguistics
UCL Institute of Education, University College London

Dr. Talia Isaacs
PhD Research Supervisor

UCL Centre for Applied Linguistics
UCL Institute of Education, University College London
Rm. 618b, 20 Bedford Way, London

Thank you for reading this information sheet and for considering taking part in this research study.

Appendix 8: Consent form for participants

CONSENT FORM FOR EXAMINERS PARTICIPATING IN APTIS RESEARCH

Please complete this form after you have read the Information Sheet about the research.

Title of Study: Communicative adequacy in the Aptis test

Department: Institute of Education, Department of Culture, Communication & Media

Name and Contact Details of the Researcher: Judith Fairbairn,

Name and Contact Details of the Supervisor: Dr. Talia Isaacs,

Name and Contact Details of the UCL Data Protection Officer:

This study has been approved by the UCL Research Ethics Committee: Project ID number: _____

Thank you for considering taking part in this research. The person organising the research must explain the project to you before you agree to take part. If you have any questions arising from the Participation Information Sheet for Aptis Examiners, please ask the Researcher before you decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time.

I confirm that I understand that by ticking/initialling each box below I am consenting to this element of the study. I understand that it will be assumed that unticked/initialled boxes means that I DO NOT consent to that part of the study. I understand that by not giving consent for any one element that I may be deemed ineligible for the study.

	Tick Box

1.	I confirm that I have read and understood the Participant Information Sheet for Aptis Examiners for the above study. I have had an opportunity to consider the information and what will be expected of me. I have also had the opportunity to ask questions which have been answered to my satisfaction.	
2.	I would like to take part in reviewing 45 Aptis Speaking Task 4 (B2) responses, answering questions about the response and writing explanations for your decisions (three hours).	
3.	I would like to take part in the two-hour email exchange interview.	
4.	I understand that I will be able to withdraw my data at any point up to 4 weeks after the interview.	
5.	I consent to the processing of my personal information (questionnaire, decisions and explanations and interview data) for the purposes explained to me. I understand that such information will be handled in accordance with all applicable data protection legislation.	
6.	All data will be held on a secure British Council server and password protected and only the Researcher will have access to the data. I understand that all personal information will remain confidential and that all efforts will be made to ensure I cannot be identified in any publications. I understand that my data gathered in this study will be stored anonymously and securely.	
7.	I understand that my information may be subject to review by responsible individuals from the University or British Council for monitoring and audit purposes.	
8.	I understand that my participation is voluntary and that I am free to withdraw at any time without giving a reason. I understand that if I decide to withdraw, any personal data I have provided up to that point will be deleted unless I agree otherwise.	
9.	I understand the potential risks of participating and the support that will be available to me.	
10.	I understand the direct/indirect benefits of participating.	

11.	I understand that the data will not be made available to any commercial organisations but is solely the responsibility of the Researcher undertaking this study.	
12.	I understand that I will be compensated for the portion of time spent in the study.	
13.	I understand that I will not benefit financially from this study or from any possible outcome it may result in in the future.	
14.	I agree that my anonymised research data may be used by others for future research. No one will be able to identify you when this data is shared.	
15.	I understand that the information I have submitted will appear in an unpublished PhD dissertation and I wish to receive a copy of it. **Yes/No**	
16.	I consent to my interview being audio recorded and understand that the recordings will be destroyed within on month after the PhD has been completed and will not be shared in public (e.g. at conferences).	
17.	I confirm that I understand the inclusion criteria (how examiners will be selected to participate in the study) as detailed in the Information Sheet (#3 Why have I been chosen).	
18.	I am aware of who I should contact if I wish to lodge a complaint.	
19.	I voluntarily agree to take part in this study.	
20.	I would be happy for the anonymised data to be archived on the British Council server for five years after the PhD has been completed. This data will not include the interview recording, which will be deleted one month after the PhD has been completed. I understand that other authenticated researchers will have access to my anonymised data.	

If you would like your contact details to be retained so that you can be contacted in the future by UCL or British Council researchers who would like

to invite you to participate in follow up studies to this project, or in future studies of a similar nature, please tick the appropriate box below.

Yes, I would be happy to be contacted in this way for follow up research	
No, I would not like to be contacted for follow up research	

_____ _____ _____

Name of participant Date Signature

Appendix 9: Participant background questionnaire

Thank you for filling in this background questionnaire. Once I have received your responses, I will send you the Speaking Task 4 responses to review within 7 days.

1. Participant code (see email for your code): _____

2. Please indicate which of the following English teaching / applied linguistics qualifications you have (select as many as appropriate).

   - ☐ CELTA
   - ☐ DELTA
   - ☐ Graduate Certificate in TESOL/TEFL/TESL
   - ☐ Graduate Diploma in TESOL/TEFL/TESL
   - ☐ Master's degree in Applied linguistics/TESOL/TESL/TEFL/Language Testing
   - ☐ Other (please specify)

3. Please indicate the approximate amount of teaching experience you have.

   - ☐ 1-3 years
   - ☐ 4-5 years
   - ☐ 6-10 years
   - ☐ More than 10 years

4. Do you have experience rating any other English language tests (e.g., IELTS, etc)? _____

5. Please list the L1 backgrounds of the learners you have experience teaching.

   _____

** In addition, in the initial recruitment email I asked them how many years they had marked Aptis tests

Appendix 10: Data collection email instructions

Dear [Rater],

Thank you for the questionnaire.

The next step is to **fill in the Excel perceptions spreadsheet** (attached). I will send a link to the speaking responses and task prompts in a separate email.

Listen to each Speaking Task 4 recording and then decide if **each of the questions in the prompt** is:

- on-topic
- off-topic
- borderline/ambiguously on-topic (you are not sure if it is on- or off-topic).

There is a drop-down list to choose one of the three options. There are 45 responses and 3 questions in each prompt (total of 135 decisions).

Write the reasons for each of the 135 decisions. Why did you decide it was on-topic? How did you make this decision?

Put the score you would assign to this test-taker in column H.

Once you have completed the spreadsheet, send it back to me by email.

Once I receive your spreadsheet, I will contact you within 14 days with the email interview questions. We will have an email exchange so that I am clear about your decisions and to discuss your perceptions in more depth.

I hope you find this experience interesting. No decision is correct / incorrect and no information will be shared with anyone else. I am genuinely very interested in how you make these decisions and I look forward to discussing your perceptions about on-topic responses.

If anything is unclear, please do not hesitate to contact me.

Best wishes,

Judith

Appendix 11: Sample for pilot studies

Regions	Region Sample	Countries	Country Sample	Pilot 1 Version 1 Score 3	Pilot 1 Version 1 Score 4	Pilot 2 Version 2 Score 3	Pilot 2 Version 2 Score 4	Pilot 3 Version 3 Score 3	Pilot 3 Version 3 Score 4
Middle East and North Africa	22	Algeria	**9**	2	2	1	2	1	1
		Egypt	**9**	2	2	2	0	2	1
		Jordan	**4**	0	0	2	0	2	0
South Asia	35	Bangladesh	**4**	2	2	0	0	0	0
		India	**12**	2	2	2	2	2	2
		Pakistan	**7**	1	2	2	0	1	1
		Sri Lanka	**12**	2	2	2	2	2	2
Americas	40	Brazil	**10**	2	2	0	2	2	2
		Chile	**10**	2	2	0	2	2	2
		Colombia	**11**	1	2	2	2	2	2
		Mexico	**9**	2	2	2	1	1	1
East Asia	54	China	**10**	2	2	2	0	2	2
		Hong Kong	**11**	2	2	2	2	2	1
		Indonesia	**12**	2	2	2	2	2	2
		S. Korea	**12**	2	2	2	2	2	2
		Malaysia	**9**	0	2	2	2	2	1
Europe	24	Poland	**12**	2	2	2	2	2	2
		Spain	**12**	2	2	2	2	2	2
		Total	**175**	**30**	**34**	**29**	**25**	**31**	**26**

Appendix 12: Semi-structured email exchange – email #1

Dear [Rater],

Thanks for your spreadsheet. Your comments were exactly what I was looking for.

This email is to start a discussion about your views on marking Aptis Speaking Task 4. There are no correct or incorrect answers and I am sincerely interested in the decisions you make when marking this task. It is ok to change your decisions or the reasons for your decisions.

I've chosen XX responses to explore in more depth and XX responses to clarify. Please re-play the responses, review your decisions and comments and answer the questions below. Just reply to this email and answer as fully as you can. I will email back if anything is unclear and also follow-up with additional questions once you have returned this email.

- What is your marking strategy? How do you normally go about marking Speaking Task 4? (This question was asked to all raters)
- Specific questions to each rater based on a content analysis of trends in the decision-making. For example:
    a. Why did you mark ___ response at ___ level?
    b. Why did you think ____ response was off-topic / on-topic?
    c. How important is it to you that a response is properly sign-posted?
    d. How important is it to you that a response is answered in a linear manner?

e. How important is it to you that a response is integrated?

f. How important is it to you that a response has an anecdote?

g. How important do you think it is for the response to be on-topic?

h. How generalisable do you think your on-topic decision is for rating other tests? Would you use this same decision-making process when marking other tests?

i. Would you ever mark a person down if they talked about feelings in the third person rather than their own personal feelings? Why or why not?

j. Which elements of linguistic ability do you think have the biggest effect on this response being on-topic or off-topic?

Thanks for any further information you can give me. It's all very interesting and useful.

Best wishes,

Judith

Appendix 13: Semi-structured email exchange – final email

Dear [Rater],

Thank you so much for the additional information. Your comments are really great.

Once I have collected all the data I may follow-up with a focus group or skype interview. Let me know if you would still like to participate.

Also, as a final question: Was there anything you felt really strongly about? One of the down-sides of email interviews is that it can be difficult to understand emotion so if there is anything where you felt any type of emotion when marking the on-topic criteria [on-topic rating requirements], please let me know. It might help me prioritise decisions as important or not so important.

I will let [Rater Manager] know to add this work into your [monthly] payment.

Thanks again for your work.

Best wishes,

Judith

Appendix 14: Example of email exchange questions

Response #	Researcher notes	Prompt 1	Prompt 1 perception	Prompt 2	Prompt 2 perception	Prompt 3	Prompt 3 perception	Aptis Score	Email exchange questions
V1-117	Why a 3 if on-topic; What is the exact quote for P2? Conflicts with others.	On-topic	played foot ball as outdoor activity'	On-topic	I feel good doing this activity'	On-topic	gives relevant examples	3	**Question to rater:** You have marked each section on-topic but scored the test taker a 3. Why not a 4? **Rater response:** Although the TT's responses are all on topic, she dedicates most of the time to answering the third question and very little time to the first. Plus, there is a high frequency of small errors in grammar and at times the pronunciation is impeding. For these reasons I penalised her by taking away one mark. I had placed her at four for complex structures and overall performance in answering question 3. **Question to rater:** Have a look at your comment for Prompt 2. It's in the present tense yet the prompt is eliciting language in the past tense. Did the tense of this response impact on your decision? **Rater response:** No, the tense in this particular sentence did not influence my decision because I understood it as a continuation from her childhood experience as to say 'it always has made me feel good to play football', including past and present.
V1-125	B1 - why? Is language B1? P1 - why is the experience of riding not on-topic? Is the benefits of sport not a reason why it's important for children to do outdoor activities?	Borderline	TT focuses on the experience of riding	Borderline	continues answer from the riding experience which may not necessarily be from childhood	Borderline	TT bases answer on the benefits of the sport	2	**Question to rater:** 1. How did you decide on a mark of 2? You have three borderline on/off-topic decisions and a final score of a high B1. How did you use the three borderline decisions to make this decision? **Rater response:** In my mind I started from a potential mark of 4 then the mark went to 3 because of the number of the pausing and lack of -ed; then I calculated that three slightly off topic sentences are equal to one fully on topic. So, instead of 3 I penalised to 2. **Question to rater:** 2. You mention that the experience of riding is borderline on/off-topic. Why? **Rater response:** After listening again, I can see TT implicitly refers to his childhood. When I had listened while filling in the spreadsheet, in my mind I thought the TT was being too general and may not necessarily be talking about his childhood—he could have been talking about anytime—when he was in high school, for example. For this same reason, I thought all the answers were not specifically for childhood and children. Now that I believe it is implicit, I would award a 4 despite his pausing because he does demonstrate enough B2 features.

Appendix 15: Qualitative analysis codes

Prompt 1:

	Perception code	Prompt 1: Tell me about a time... Examiner perceptions of test taker responses * Why did you make this on-topic decision? * How did you make this on-topic decision?	Generic Criteria Codes: Amount of response	Clarity of response	Picture	Words in prompt	Example	Quote	Format	Pre-prepared response	No evidence	Off topic	Task Prompt 1: Multiple activities, journeys or buildings	Task Versions: Not an outdoor activity (V1)	Not a journey (V2)	Not a tall building (V3)	Rating Scale: Verb tenses	Vocabulary	Other language or overall language
3	V1-107-06	Describes drawing. Not a typical outdoor activity, but answers the question as it is an activity that he did outside, drawing natural scenery.					Y							Y					
4	V1-112-07	Although the candidate took the photo as her starting point and began by talking about cycling, she soon widened this to talk more about other outdoor activities.	Y		Y								Y						

Prompt 2:

	Perception code	Prompt 2: How did you feel about... Examiner perceptions of test taker responses * Why did you make this on-topic decision? * How did you make this on-topic decision?	Generic Criteria Codes: Amount of response	Clarity of response	Picture	Words in prompt	Example	Quote	Format	Pre-prepared response	No evidence	Off topic	Prompt 2: Does not describe a feeling	Feeling is incorrect (not a feeling)	Task Versions: Not an outdoor activity (V1)	Not a journey (V2)	Not about cities and tall buildings (V3)	Rating Scale: Verb tenses	Vocabulary	Other language or overall language
3	V1-112-06	Not much detail but says she 'loved cycling' and answers to Q1 & Q3 implies her enthusiasm	Y	Y			Y	Y							Y					
4	V1-113-07	Didn't address this, however the sound recording was very poor and I couldn't hear anything after about 1 minute 12 seconds.										Y								

Prompt 3

	Perception code	Prompt 3: Abstract question Examiner perceptions of test taker responses * Why did you make this on-topic decision? * How did you make this on-topic decision?	Generic Criteria Codes: Amount of response	Clarity of response	Picture	Words in prompt	Example	Quote	Format	Pre-prepared response	No evidence	Off topic	Prompt 3: Abstract question - does not generalise	Rating Scale: Verb tenses	Vocabulary	Other language or overall language
3	V1-113-06	Addresses this fairly well; gives 2 clear reasons and attempts to give a third. Unfortunately, she becomes tongue-tied and stops well before time is up.	Y	Y					Y							Y
4	V1-117-07	This comprised the main bulk of the response but was very repetitive and she ran out of steam at the end and the last 20 seconds were in silence.	Y	Y					Y							Y

Appendix 16: Information sheet for the second coder

Qualitative analysis research questions:

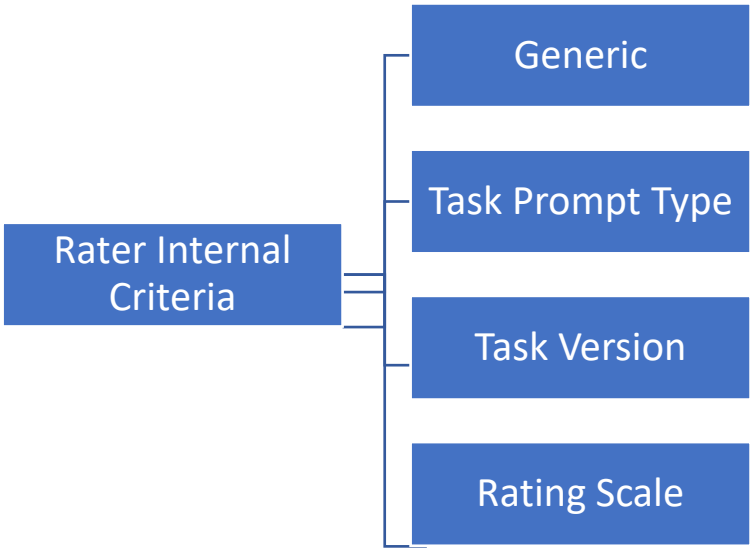
1. What internal criteria do raters report using to decide if a speaking response is on-topic?

    1.1. What generic criteria do raters report using that are common across all prompts and versions?

    1.2. How do B2-level Aptis speaking task prompt types affect raters' decision-making?

    1.3. How do the task versions from different forms of the test affect raters' decision-making?

    1.4. How do the internal criteria that raters reportedly use reflect the criteria in the rating scale descriptors?



Codes:

- Raters answered the following questions:
    - Why did you make this on-topic decision?
    - How did you make this on-topic decision?

Generic codes: internal rater perceptions that are generic across all prompts and versions

- Amount of test taker response (e.g., brief, not enough detail, depth, detailed, at length)
- Clarity of test taker response (e.g., clear, clearly, not clear, implied, implicit, explicit)
- Picture mentioned
- Words in prompt used as evidence
- Examples from test taker response used for evidence
- Quotes from test taker response used for evidence
- Format of test taker response mentioned (e.g., integrated response, well structured)
- Pre-prepared test taker response
- Test taker response is off topic

Task Prompt Type codes: internal rater perceptions specific to the prompt type:

Prompt 1: Tell me about a time…

- Multiple (not one) activities, journeys or buildings mentioned by rater

Prompt 2: How did you feel about…

- Description of feeling; types of feelings
- Rater does not describe a feeling
- Feeling is incorrect, i.e., not a feeling

Prompt 3: abstract question

- Does not generalise (e.g., talks about a personal experience)

Task Version codes: internal rater perceptions specific to the task version:

Version 1: Outdoor activities in childhood

- Not an outdoor activity

Version 2: Long journeys

- Trip or holiday or destination impacts on decision (i.e., not a journey)
- Length or distance of journey impacts on decision

[No Version 3]

Rating Scale Codes: internal rater perception that mentioned language (other Aptis rating scale criteria)

- Verb tenses (talking in the wrong tense)
- Vocabulary or lexis
- Other language (e.g., grammar - not verb tenses, pronunciation, fluency, discourse markers, coherence, cohesion) or overall language

**I also included the task versions (Appendix 2) in the email to the second coder.